NISS

Multiple Hypothesis Testing: A Review

Juliet Popper Shaffer

Technical Report Number 23
September, 1994

National Institute of Statistical Sciences
19 T. W. Alexander Drive
PO Box 14006
Research Triangle Park, NC 27709-4006
www.niss.org

MULTIPLE HYPOTHESIS TESTING: A REVIEW *

Juliet Popper Shaffer

Department of Statistics, University of California, Berkeley, California 94720

ABSTRACT

This is a review of multiple hypothesis testing methods that control in some overall way the probabilities of rejecting true null hypotheses. Two types of procedures are considered: (a) methods based on ordered p-values, and (b) methods for comparing normally-distributed means. Recent results are emphasized.

A condensed version of this review will appear in the 1995 Annual Review of Psychology.

KEY WORDS: multiple comparisons, simultaneous testing, p-values, closed test procedures, pairwise comparisons

* Research supported in part through the National Institute of Statistical Sciences by the National Science Foundation Grant No. RED-9350005. Thanks to Yosef Hochberg, Lyle V. Jones, Erich L. Lehmann, Barbara A. Mellers, Seth D. Roberts, and Valerie S.L. Williams for helpful comments and suggestions.

CONTENTS

INTRODUCTION

Examples

A Brief History

Books

Scope of this review

ORGANIZING CONCEPTS

Primary Hypotheses, Closure, Hierarchical Sets, Minimal Hypotheses

Families

Type I Error Control and Power

P-values and Adjusted P-values

Closed Test Procedures

METHODS BASED ON ORDERED P-VALUES

Methods Based on the First-order Bonferroni Inequality

Methods Based on the Simes Equality

Modifications for Logically Related Hypotheses

Methods Controlling the FDR

COMPARING NORMALLY-DISTRIBUTED MEANS

OTHER ISSUES

Tests vs. Confidence Intervals

Directional vs. Nondirectional Inference

Robustness

Others

CONCLUSION

REFERENCES

INTRODUCTION

Multiple testing refers to the testing of more than one hypothesis at a time. It is a subfield of the broader field of multiple inference, or simultaneous inference, which includes multiple estimation as well as testing. The term 'multiple comparisons' has come to be used synonymously with 'simultaneous inference', even when the inferences do not deal with comparisons. It will be used in this broader sense throughout this review.

In general, in testing any single hypothesis, conclusions based on statistical evidence are uncertain. We typically specify an acceptable maximum probability of rejecting the null hypothesis when it is true, or committing a Type I error, and base the conclusion on the value of a statistic meeting this specification, preferably one with high power. When many hypotheses are tested, and each test has a specified Type I error probability, the probability that at least some Type I errors are committed increases, often sharply, with the number of hypotheses. This may have serious consequences if the set of conclusions must be evaluated as a whole. Numerous approaches to this issue and many methods for resolving it have been proposed. No one solution will be acceptable for all situations.

Examples

In order to focus the discussion, several examples are given below illustrating different types of multiple testing problems. More examples can be found in the books described later and in Diaconis (1985).

A. Subpopulations: A historical example. Cournot (1843) described vividly the multiple testing problem resulting from the exploration of effects within different subpopulations of an overall population. In his words (translated from the French): "...it is clear that nothing limits ... the number of features according to which one can distribute

[natural events or social facts] into several groups or distinct categories." As an example he mentions investigating the chance of a male birth. "One could distinguish first of all legitimate births from those occurring out of wedlock, ... one can also classify births according to birth order, according to the age, profession, wealth, or religion of the parents... ." He goes on to point out that as one increases the number of such "cuts" (of the material into two categories) it becomes more and more likely that by pure chance for at least one pair of opposing categories the observed difference will be significant. "As a result... the probability that an observed deviation can not be attributed to the vagaries of chance takes on very different values depending on whether one has tried a more or less large number of cuts before having hit on the observed deviation. ... Usually these attempts through which the experimenter passed don't leave any traces; the public will only know the result that has been found worth pointing out; and as a consequence, someone unfamiliar with the attempts which have led to this result completely lacks a clear rule for deciding whether the result can or can not be attributed to chance." (See Stigler 1986, for further discussion of the historical context.)

While these issues are still serious and far from being solved (see e.g. Nowak 1994), multiple comparison methods provide a means for approaching such problems. It is vital to obtain solutions--in medicine, in education, in all social and behavioral sciences--where the effects of experimental treatments or environmental events may vary over subpopulations defined in many ways. (See Shafer & Olkin 1983 for work closely related to this issue.)

B. Multiple outcomes: Drug screening for carcinogenic effects. Many potentially-useful drugs are initially screened for carcinogenic effects in animal studies. There are about 15 major cancer sites that are monitored; correlations between effects at different sites are positive but rather low. Thus, if the effects at each site are tested individually at typical Type I error levels, the probability that one or more sites will

show apparent carcinogenic activity may approach .50. Any carcinogenic effect will eliminate the use of a drug, so many promising candidates could be eliminated using such a procedure. On the other hand, it is vital to detect such effects if they are present. See e.g. Heyse & Rom (1988), Rom (1992).

C. Interim Tests. In clinical trials of new treatments, early results may indicate such marked positive effects of an innovation that it appears unethical to continue the trials in which some patients are receiving alternative medical interventions. Thus, many trials allow for repeated testing. If no allowance is made for the effect of multiple testing, an unacceptably high level of Type I error may result. This is a very active area of current research, with potential applications in many fields; for reviews see Armitage (1993), DeMets (1987), Geller & Pocock (1987), Jennison & Turnbull (1990).

It should be pointed out that although a trial may be designed to compare a single treatment with a control group, there are always multiple outcomes to consider, since side effects of any new treatment must be monitored. Thus, the multiplicity due to multiple outcomes, as described in Example B, and to interim testing, must be considered jointly.

D. Large surveys and observational studies. In large social science surveys, thousands of variables are investigated, and participants are grouped in myriad ways. The results of these surveys are often widely publicized and have potentially large effects on legislation, monetary disbursements, public behavior, etc. Thus, it is important to analyze results in such a way as to minimize misleading conclusions. Some type of multiple error control is needed, but it is clearly impractical, if not impossible, to control errors at a small level over the entire set of potential comparisons. For discussion of these issues see Ahmed (1991).

Some critics maintain that results of large surveys should be simply tabulated and presented without any conclusions as to which are reliable, and that such decisions should be left up to the individuals reading the reports. But statistically unsophisticated readers are likely to draw far too many unsubstantiated conclusions if such policies were to be adopted.

E. Apparent clusters in space and/or time. There is constant publicity concerning apparently high rates of cancer in specific localities, apparent crime surges during particular time periods in some places, etc. Since some such events would be expected by chance, methods are needed for deciding whether particular occurrences should be considered random, or evidence of underlying problems. Of course, investigations of these apparent clusters involve consideration of many associated variables that could conceivably have causal connections to the target events. But multiplicity issues remain, and far too often the outcomes of such investigations are inconclusive. For methods used in this area see Williams (1984).

F. Factorial designs. The standard textbook presentation of multiple comparison issues is in the context of a one-factor investigation, where there is evidence from an overall test that the means of the dependent variable for the different levels of a factor are not all equal, and more specific inferences are desired to delineate which means are different from which others. Here, in distinction to many of the examples above, the family of inferences for which error control is desired is usually clearly specified, and is often relatively small. On the other hand, in multifactorial studies, the situation is less clear. The typical approach is to treat the main effects of each factor as a separate family for purposes of error control, although both Tukey (1953) and Hartley (1955) gave examples of 2 x 2 x 2 factorial designs in which they treated all seven main effect and interaction tests as a single family. The probability of finding some significances may be very large if each of many main effect and interaction tests is carried out at a

conventional level in a multifactor design. Furthermore, it is important in many studies to assess the effects of a particular factor, say A, separately at each level of other factors. thus bringing in another layer of multiplicity (see Shaffer 1991).

In spite of the importance of factorial designs and their wide use, there has been relatively little discussion of these issues in the literature.

A Brief History

As noted in Example A, Cournot (1843) clearly recognized the problems involved in multiple inference, but considered them insoluble. Although there were a few isolated earlier relevant publications, sustained statistical attacks on the problems began in the late 1940s and early 1950s. Papers by Mosteller (1948) and Nair (1948) dealt with extreme value problems, while a more comprehensive approach was published by Tukey (1949). Duncan (1951) treated multiple range tests. Related work on ranking and selection was published by Paulson (1949) and Bechhofer (1952). Scheffé (1953) introduced his well-known procedures, and work by Roy and Bose in that same year (1953) developed another simultaneous confidence interval approach. Also in that year, a book-length unpublished manuscript by Tukey presented a general framework covering a number of aspects of multiple inference. This manuscript remained unpublished until recently, when it was reprinted in full (Braun 1994). In the later 1950s a decision-theoretic approach was developed by Lehmann (1957a,b), and a Bayesian decision-theoretic approach was developed shortly afterwards by Duncan (1961). For additional historical material see Tukey (1953), Harter (1980), Miller (1981), Hochberg & Tamhane (1987), and Shaffer (1988).

Books

The first published book on the subject was Miller (1966); it was reissued in 1981 (Miller 1981), unchanged except for the addition of a 1977 review article (Miller 1977).

For some time there were no additional book-length treatments, except in the ranking and selection area. Then, in 1986, a series of books began to appear. A brief synopsis of each of these newer volumes is given below.

MULTIPLE COMPARISONS (Klockars & Sax, 1986; 87 p.) This is an introductory treatment oriented towards social scientists. It treats general issues, and specific methods for normally-distributed observations in one-way randomized designs, with some consideration of multifactor completely randomized designs.

MULTIPLE COMPARISON PROCEDURES (Hochberg & Tamhane 1987; xxii, 450 p.) This is a comprehensive reference work with good background information on general approaches, theoretical issues, and useful probability distributions and inequalities, as well as extensive coverage of methodology specific to linear models and a survey of methods appropriate to categorical data and contingency table analysis. It does not deal specifically with multivariate analysis, although the general material is applicable in that area. For reviews, see Littell (1989) and Peritz (1989).

MULTIPLE HYPOTHESENPRÜFUNG (MULTIPLE HYPOTHESES TESTING (Bauer, Hommel, & Sonnemann 1988; ix, 234 p.) This is an edited bilingual volume containing papers delivered at a two-day symposium on "Multiple Hypotheses Testing" held November 6-7, 1987 in the Federal Republic of Germany. The preface and summaries of all papers are in both German and English; of the 18 papers, ten are in German and eight in English. There are a number of stimulating theoretical discussions, and a variety of interesting probability inequalities are presented and associated test procedures described. Most papers are theoretical, but there are also some simulation studies comparing the powers of alternative methods. For reviews see Läuter (1990) and Holm (1990).

MULTIPLE COMPARISONS FOR RESEARCHERS (Toothaker 1991; viii, 168 p.) and MULTIPLE COMPARISON PROCEDURES (Toothaker 1993; viii, 96 p.). The

1991 volume is for applied researchers; the 1993 volume is a condensed and simplified version. They constitute an introduction to the subject containing a minimum of statistical theory, and a number of numerical illustrations of the methods covered, which are mainly those designed for the one-way ANOVA model, with a brief discussion of extensions to two-way models. There are nontechnical discussions of issues connected to the use of the methods. The material is a subset, at a much less technical level, of that covered in the Hochberg & Tamhane book. For reviews of the 1991 volume, see Gaffan (1992) and Tatsuoka (1992).

MULTIPLE COMPARISONS, SELECTION, AND APPLICATIONS IN BIOMETRY (Hoppe 1993; xii, 558 p.). Subtitled "A Festschrift in Honor of Charles W. Dunnett," this is an edited volume containing papers on multiple comparisons, selection, and specific applications to biometry. The papers cover a large variety of situations, and vary greatly in difficulty and in theoretical vs. applied emphasis. For a review, see Ziegel (1994).

RESAMPLING-BASED MULTIPLE TESTING (Westfall & Young 1993; xvii, 340 p.) This volume describes a comprehensive approach to the use of bootstrap and permutation methods with univariate and multivariate data. There are valuable general discussions of multiple comparisons (with a useful discussion of their utility in comparison to meta- analysis) and of bootstrap methods, and a variety of specific applications are analyzed in detail. Many examples of the type described above are considered, with discussions more extensive than those possible in this review. The book advocates the use of adjusted p-values (to be described in a later section of this report) for interpretation. The resampling methods advocated are highly computer-intensive. Detailed discussion of computer implementation is provided. For reviews see Chaubey (1993) and Booth (1994).

THE COLLECTED WORKS OF JOHN W. TUKEY, VOLUME VIII: MULTIPLE COMPARISONS: 1948-1983. (Braun 1994; lxi, 475 p., i10) This volume contains contributions of Tukey from 1948 through 1983 with an almost equal mixture of published and unpublished works. As noted above, the volume contains the first published version of the 1953 book-length manuscript "The Problem of Multiple Comparisons," still much worth reading for its coverage of general concepts, although of course many of the specific methods discussed have been superseded. (Tukey notes in the Foreword to the Volume that this manuscript was unpublished due to "the only piece of bad advice I ever had from Walter Shewhart! He told me it was unwise to put a book out until one was sure that it contained the last word of one's thinking.")

MULTIPLE COMPARISONS: THEORY AND METHODS (Hsu 1996) This volume, to be published in 1996, is organized around a classification of procedures in terms of type and strength of permissible inference. It deals primarily with linear models and normally-distributed errors. It has a nice discussion of relevant probability inequalities, many elegant proofs, and illuminating geometric explanations. The theoretical discussions are supplemented with detailed computer implementation guidelines.

Scope of This Review

The field of multiple testing is too broad to be covered in its entirety in a review of this length. In consequence, apologies are due to many active researchers whose valuable contributions may not be acknowledged. Most attention will be devoted to two types of methods, (1) Methods based on ordered p-values, and (2) Comparisons among normally-distributed means.

ORGANIZING CONCEPTS

Primary Hypotheses, Closure, Hierarchical Sets, Minimal Hypotheses

Assume some set of null hypotheses of primary interest to be tested. Sometimes the number of hypotheses in the set is infinite (e.g. hypothesized values of all linear contrasts among a set of population means), although in most practical applications it is finite (e.g. values of all pairwise contrasts among a set of population means). It is assumed that there is a set of observations with joint distribution depending on some parameters, and that the hypotheses specify limits on the values of those parameters. I'll give examples using a primary set based on differences among the means $\mu_1,\mu_2,...,\mu_m$ of m populations, although the concepts apply in general. Let δ_{ij} be the difference $\mu_i-\mu_j$; let δ_{ijk} be the set of differences among the means μ_i, μ_j, and μ_k, etc. The hypotheses will be of the form $H_{ijk...}$: $\delta_{ijk...}=0$, indicating that all subscripted means are equal; e.g. H_{1234} is the hypothesis $\mu_1=\mu_2=\mu_3=\mu_4$. Note that the primary set need not consist of the individual pairwise hypotheses H_{ij}. If m = 4, it may, for example, be the set $H_{12}, H_{123}, H_{1234}$, etc., which would signify a lack of interest in including inference concerning some of the pairwise differences (e.g. H_{23}) and therefore no need to control errors with respect to those differences.

The *closure* of the set is the collection consisting of the original set together with all distinct hypotheses formed by intersections of hypotheses in the set; such a collection is called a *closed set.* For example, an intersection of the hypotheses H_{ij} and H_{ik} is the hypothesis H_{ijk}: $\mu_i=\mu_j=\mu_k$. Any hypotheses included in an intersection are called components of the intersection hypothesis. Technically, a hypothesis is a component of itself; any other component is called a proper component. In the example above, the proper components of H_{ijk} are H_{ij}, H_{ik} and, if it is included in the set of primary interest, also H_{jk}, since its intersection with either of H_{ij}, H_{ik} also gives H_{ijk}. Note that the truth of a hypothesis implies the truth of all its proper components.

Any set of hypotheses in which some are proper components of others will be called a *hierarchical set.* (The term is sometimes used in a more limited way, but this definition will be adopted here.) A closed set (with more than one hypothesis) is therefore a hierarchical set. In a closed set, the top of the hierarchy is the intersection of all hypotheses: in the examples above it is the hypothesis $H_{12\ldots m}$, or $\mu_1=\mu_2=\cdots=\mu_m$. The set of hypotheses that have no proper components comprise the lowest level of the hierarchy; these are called the *minimal hypotheses* (Gabriel 1969). Equivalently, a minimal hypothesis is one that does not imply the truth of any other hypothesis in the set. For example, if all the hypotheses state that there are no differences among sets of means, and the set of primary interest includes the hypotheses H_{ij} for all $i\neq j=1,\ldots,m$, these pairwise equality hypotheses are the only minimal hypotheses.

Families

The first and perhaps most crucial decision is what set of hypotheses to treat as a family, that is, as the set for which significance statements will be considered and errors controlled jointly. In some of the early multiple comparisons literature, e.g. Ryan (1959, 1960), the term 'experiment' rather than 'family' was used in referring to error control. The implication was that an explicit control of error should apply jointly to all potential inferences resulting from a total experiment. Implicitly, attention was directed to relatively small and limited experiments. As a dramatic contrast, consider large surveys as described in Example D. Here, because of the inverse relationship between control of Type I errors and power, it is unreasonable if not impossible to consider methods controlling the error rate at a conventional level, or indeed any level, over all potential inferences from such surveys. An intermediate case is a multifactorial study, where as noted in Example F, it frequently seems unwise from the point of view of power to control error over all inferences. The term "family" was introduced by Tukey

(1952, 1953), and allows flexibility in defining the set of actual or potential inferences to be considered jointly in bounding some function of Type I error. Miller (1981), Diaconis (1985), Hochberg & Tamhane (1987) and others discuss the issues involved in deciding on a family. They stress the difference between exploratory research, involving data snooping, and confirmatory research, where the questions of interest are well defined. Although a less stringent criterion might be adopted in exploratory studies, nonetheless some allowance for multiplicity in such studies will prevent wasteful followup of illusory leads; consider especially examples A, D, and E. The most comprehensive discussion of these issues is in the book by Westfall & Young (1993), who give explicit advice on methods of approaching complex experimental studies.

Since a study can be used for different purposes, the results may have to be considered under several different family configurations. This issue came up in reporting state and other geographical comparisons in the National Assessment of Educational Progress Trial State Assessment, a part of a large national survey designed to compare geographical jurisdictions on the educational achievements of their school-age children. In a recent national report, each of the 780 pairwise differences on a measure of achievement among the 40 jurisdictions involved (states, territories, District of Columbia) was tested for significance at level .05/780 in order to control Type I errors for that family. However, from the point of view of a single jurisdiction, the family of interest is the 39 comparisons of itself with each of the others, so it would be reasonable to test those differences each at level .05/39, in which case some differences would be declared significant that were not so designated in the national report. For a discussion of this example and other issues in the context of large surveys, see Ahmed (1991).

Type I Error Control and Power

In testing a single hypothesis, the probability of a Type I error, i.e. of rejecting the null

hypothesis when it is true, is usually controlled at some designated level α. The choice of α should be governed by considerations of the costs of rejecting a true hypothesis as compared with those of accepting a false hypothesis. Because of the difficulty of quantifying these costs and the subjectivity involved, α is usually set at some conventional level, often .05. A variety of generalizations to the multiple testing situation are possible.

Some multiple comparison methods control the Type I error rate only when all null hypotheses in the family are true. Others control this error rate for any combination of true and false hypotheses. Hochberg & Tamhane (1987) refer to these as weak control and strong control, respectively. Examples of methods with only weak error control are the Fisher protected least significant difference (LSD) procedure, the Newman-Keuls procedure, and some nonparametric procedures. (For discussion of the first two, see Keselman, Keselman, & Games 1991. For the latter see Fligner 1984.) The multiple comparison literature has been confusing because the distinction between weak and strong control is often ignored. In fact, weak error rate control without other safeguards is unsatisfactory; for example, such control is achieved in a one-way layout satisfying standard linear model conditions if a significant F test results in the decision that all differences among means are significant without further testing. This review will concentrate on procedures with strong control of the error rate.

Several different error rates have been considered in the multiple testing literature. The major ones are listed below.

The *error rate per hypothesis (PCE)* (usually called PCE, for per-comparison error rate, although the hypotheses need not be restricted to comparisons) is defined for each hypothesis as the probability of Type I error or, when the number of hypotheses is finite, the average PCE can be defined as the expected value of (Number of false rejections/ Number of hypotheses), where a false rejection means the rejection of a true

hypothesis.

The *error rate per family (PFE)* is defined as the expected number of false rejections in the family. Note that this error rate doesn't apply if the family size is infinite.

The *error rate familywise or familywise error rate (FWE)* is defined as the probability of at least one error in the family.

A fourth type of error rate, the false discovery rate (FDR) will be described below. First, to make the three definitions above clearer, consider what they imply in a simple example in which each of n hypotheses $H_1, \cdots H_n$ is tested individually at a level α_i, and the decision on each is based solely on that test. (Procedures of this type are called *single-stage;* other procedures have a more complicated structure.) If all the hypotheses are true, the average PCE equals the average of the α_i, the PFE equals the sum of the α_i, and the FWE is a function of both the α_i and the joint distribution of the test statistics; it is between the largest α_i and the PFE.

A common misconception of the meaning of an overall error rate α applied to a family of tests is that on the average, only a proportion α of the rejected hypotheses are true ones, i.e. are falsely rejected. That this is not so can be seen by considering the case in which all the hypotheses are true; then 100% of rejected hypotheses are true, i.e. are rejected in error, in those situations in which any rejections occur. This misconception, however, suggests considering the proportion of rejected hypotheses that are falsely rejected and trying to control this proportion in some way. Letting V equal the number of false rejections (i.e. rejections of true hypotheses) and R equal the total number of rejections, the proportion of false rejections is Q = V/R. Some interesting early work related to this ratio is described in Seeger (1968), who credits the initial investigation to unpublished papers of Eklund. A recent reference is Sorić (1989), who describes a different approach to this ratio. These papers (Seeger, Eklund, and Sorić) advocated informal consideration of the ratio. A new, more formal approach is based

on the following definition.

The *false discovery rate (FDR),* in the terminology of Benjamini & Hochberg (1994a), is the expected value of Q = (Number of false significances/Number of significances).

In moving from single to multiple testing, just as error rates can be generalized in different ways, so can power. Three definitions have been common: the probability of rejecting at least one false hypothesis, the average probability of rejecting the false hypotheses, and the probability of rejecting all false hypotheses. When the family consists of pairwise mean comparisons, these have been called, respectively, any-pair power (Ramsey 1978), per-pair power (Einot & Gabriel 1975), and all-pairs power (Ramsey 1978). Ramsey (1978) showed that the difference in power between single-stage and multistage methods with strong control of the FWE is much greater for all-pairs than for any-pair or per-pair power. See also Gabriel (1978), Hochberg & Tamhane (1987), for discussion of these results.

P-values and Adjusted P-values

In testing a single hypothesis, investigators in general have moved away from simply accepting or rejecting the hypothesis toward giving the *p-value* connected with the test, i.e. the probability of observing a test statistic as extreme or more extreme in the direction of rejection as the observed value. This can be conceptualized as the level at which the hypothesis would just be rejected, and therefore both allows individuals to apply their own criteria and gives more information than merely acceptance or rejection. Extension of this concept in its full meaning to the multiple testing context is not necessarily straightforward. One generalization of the p-value for the test of a single hypothesis to the multiple context is the *adjusted p-value,* introduced by Rosenthal & Rubin (1983). Given any test procedure, the adjusted p-value corresponding to the test

of a single hypothesis H_i can be defined as the level of the entire test procedure at which H_i would just be rejected, given the values of all test statistics involved. For example, if each of n hypotheses is tested at level α/n in order to control the FWE at α (see the description of the Bonferroni procedure below), the adjusted p-value for each hypothesis H_i is np_i, where p_i is the unadjusted p-value. Application of this definition in complex multiple comparison procedures is discussed by Wright (1992), and by Westfall & Young (1993), who base their methodology on the use of such values. In addition to the generalization to the multiple testing context these values are also interpretable on the same scale as those for tests of individual hypotheses, making comparison with single hypothesis testing easier.

Closed Test Procedures

Most of the multiple comparison methods in use are designed to control the FWE. The most powerful of these methods are in the class of *closed test procedures,* described in Marcus, Peritz, & Gabriel (1976). To define this general class, assume a set of hypotheses of primary interest, add hypotheses as necessary to form the closure of this set, and recall that the closed set consists of a hierarchy of hypotheses. The closure principle is as follows: A hypothesis is rejected at level α if and only if it and every hypothesis directly above it in the hierarchy (i.e. every hypothesis that includes it in an intersection and thus implies it) is rejected at individual level α. For example, given four means, with the minimal hypotheses the six hypotheses $H_{ij}, i \neq j = 1,...,4$, the highest hypothesis in the hierarchy is H_{1234}, and no hypothesis below it can be rejected unless it (H_{1234}) is rejected at level α. Assuming it is rejected, the hypothesis H_{12} can't be rejected unless the three other hypotheses above it in the hierarchy, H_{123}, H_{124}, and the intersection hypothesis H_{12} and H_{34} (i.e. the single hypothesis $\mu_1=\mu_2$ and $\mu_3=\mu_4$) are rejected at level α, and then it (H_{12}) is rejected if its associated test statistic is

significant at that level. Any tests can be used at each of these levels, provided the choice of tests does not depend on the observed configuration of the means. The proof that closed test procedures provide strong control of the FWE involves the following simple logical argument. Consider every possible true situation, each of which can be represented as an intersection of null and alternative hypotheses. Note that only one of these possible situations can be the true one, and that under a closed testing procedure the probability of rejecting that one true configuration is $\leq\alpha$. All true null hypotheses in the primary set are contained in the intersection corresponding to the true configuration, and none of them can be rejected unless that configuration is rejected. Therefore, the probability of one or more of these true primary hypotheses being rejected is $\leq\alpha$.

METHODS BASED ON ORDERED P-VALUES

The methods discussed in this section will be defined in terms of a finite family of hypotheses H_i, i = 1, ..., n, consisting of minimal hypotheses only. It will be assumed that for each hypothesis H_i there is a corresponding test statistic T_i with a distribution that depends only on the truth or falsity of H_i. It will further be assumed that H_i is to be rejected for large values of T_i. (The T_i are absolute values for two-sided tests.) Then the (unadjusted) p-value p_i of H_i is defined as the probability that T_i is larger than or equal to t_i, where T refers to the random variable and t to its observed value. For simplicity of notation, assume the hypotheses are numbered in the order of their p-values so that $p_1 \leq p_2 \leq ... \leq p_n$, with arbitrary ordering in case of ties.

With the exception of the subsection of methods controlling the FDR, all methods in this section are intended to provide strong control of the FWE.

Methods Based on the First-order Bonferroni Inequality

The first-order Bonferroni inequality states that, given any set of events $A_1, A_2, \ldots, A_n$, the probability of their union (i.e. of the event A_1 *or* A_2 *or* $\cdots$ *or* A_n) is smaller than

or equal to the sum of their probabilities. Letting A_i stand for the rejection of H_i, i = 1,..., n, this inequality is the basis of the Bonferroni methods discussed in this section.

THE SIMPLE BONFERRONI METHOD Reject H_i if $p_i \leq \alpha_i$, where the α_i are chosen so that their sum equals α. Usually, the α_i are chosen to be equal (all equal to α/n), and the method is then called the unweighted Bonferroni method. This procedure controls the PFE to be $\leq\alpha$--exactly α if all hypotheses are true. The FWE is usually $< \alpha$.

This simple Bonferroni method is an example of a single-stage testing procedure. In single-stage procedures, control of the FWE has the consequence that the larger the number of hypotheses in the family, the smaller the average power for testing the individual hypotheses. In multistage testing procedures, in which the levels at which individual hypotheses are tested depend on decisions with respect to other hypotheses, the power for testing the individual hypotheses also decreases as the number of hypotheses in the family increases, but often to a lesser extent. The following are some multistage modifications of the Bonferroni method.

THE SEQUENTIALLY-REJECTIVE BONFERRONI METHOD (Holm 1979a) The unweighted method only will be described here; for the weighted method see Holm (1979a). This method is applied in stages as follows: At the first stage, H_1 is rejected if $p_1 \leq \alpha/n$. If H_1 is accepted, all hypotheses are accepted without further test; otherwise, H_2 is rejected if $p_2 \leq \alpha/(n-1)$. Continuing in this fashion, at any stage j, H_j is rejected if and only if all H_i have been rejected for i < j, and $p_j \leq \alpha/(n-j+1)$.

To prove that this method provides strong control of the FWE, let k be the number of hypotheses that are true, where k is some number between 0 and n. If k=n, the test at the first stage will result in a Type I error with probability $\leq\alpha$. If k = n-1, an error might occur at the first stage but will certainly occur if there is a rejection at the second stage, so again the probability of a Type I error is $\leq\alpha$ [since there are n-1 true hypotheses and none can be rejected unless at least one has an associated p-value $\leq\alpha/(n-1)$]. Similarly,

whatever the value of k, a Type I error may occur at an early stage but will certainly occur if there is a rejection at stage n-k+1, in which case the probability of a Type I error is $\leq\alpha$. Thus, the FWE is $\leq\alpha$ for every possible configuration of true and false hypotheses.

A MODIFICATION FOR INDEPENDENT AND SOME DEPENDENT STATISTICS

If test statistics are independent, the Bonferroni procedure and the Holm modification described above can be improved slightly by replacing α/k, for any k = 1,...,n, by $1-(1-\alpha)^{(1/k)}$, which is always $> \alpha/k$, although the difference is small for small values of α. As Holland & Copenhaver (1988) point out, these somewhat higher levels can also be used when the test statistics are *positive orthant dependent,* i.e. for which the joint probability that each is smaller than some individual fixed value is at least as large as if the test statistics were independent. The two-sided t statistics for pairwise comparisons of normally-distributed means in a one-way layout are positive orthant dependent, as are any symmetric two-sided test statistics for hypotheses concerning linear combinations of normally-distributed means with arbitrary positive-definite covariance structure. Holland & Copenhaver (1988) give examples of other positive orthant dependent statistics.

Methods Based on the Simes Equality

Simes (1986) proved that if a set of hypotheses $H_1,H_2,...,H_n$ are all true, and the associated test statistics are continuous and independent, then with probability $1-\alpha$, $p_i > i\,\alpha/n$ for all i = 1,...,n, where α is any number between 0 and 1. Furthermore, although Simes noted that the probability of this joint event could be smaller than $1-\alpha$ for dependent test statistics (see Hommel 1983 for a lower limit to the probability), this appeared to be true only in rather pathological cases. Simes and others (Hommel 1988, Holland 1991, Klockars & Hancock 1992, Hochberg & Blair 1994) have provided

simulation results suggesting that the probability of the joint event is larger than $1-\alpha$ for a number of types of dependence found in typical testing situations, including the important case of the usual two-sided t test statistics for all pairwise comparisons among normally- distributed treatment means.

Simes suggested that this result could be used in multiple testing but did not provide a formal procedure. As Hochberg (1988) and Hommel (1988) pointed out, on the assumption that the inequality applies in a testing situation, more powerful procedures than the sequentially rejective Bonferroni can be obtained by invoking the Simes result in combination with the closure principle. Since carrying out a full Simes-based closure procedure testing all possible hypotheses would be tedious with a large closed set, Hochberg (1988) and Hommel (1988) each give simplified, conservative methods of utilizing the Simes result.

HOCHBERG'S MULTIPLE TEST PROCEDURE Hochberg's (1988) procedure can be described as a "step-up" modification of Holm's procedure. Consider the set of primary hypotheses $H_1, \ldots, H_n$. If $p_j \leq \alpha/(n-j+1)$ for any j = 1, ..., n, reject all hypotheses H_i for $i \leq j$. In other words, if $p_n \leq \alpha$ reject all H_i; otherwise if $p_{n-1} \leq \alpha/2$, reject $H_1,\ldots,H_{n-1}$, etc.

HOMMEL'S MULTIPLE TEST PROCEDURE Hommel's (1988) procedure is more powerful than Hochberg's but more difficult to understand and apply. Let j be the largest integer for which $p_{n-j+k} > k\alpha/j$ for all k = 1, ..., j. If no such j exists, reject all hypotheses. Otherwise reject all H_i with $p_i \leq \alpha/j$.

ROM'S MODIFICATION OF HOCHBERG Rom (1990) gave slightly higher critical p-value levels that can be used with Hochberg's procedure, making it somewhat more powerful. The values must be calculated; see Rom (1990) for details and a table of values for small n.

Modifications for Logically Related Hypotheses

Shaffer (1986) pointed out that Holm's sequentially-rejective multiple test procedure can be improved when hypotheses are logically related; the same considerations apply to multistage methods based on Simes' equality. In many testing situations, it is not possible to get all combinations of true and false hypotheses. For example, if the hypotheses refer to pairwise differences among treatment means, it is impossible to have $\mu_1=\mu_2$ and $\mu_2=\mu_3$ but $\mu_1\neq\mu_3$. Using this reasoning, with four means and six possible pairwise equality null hypotheses, if all six are not true, then at most three are true, since if there are any differences at all, at least one mean must be different from the other three. Therefore, it isn't necessary to protect against error in case five hypotheses are true and one is false, for example, since this combination is impossible. Let t_j be the maximum number of hypotheses that can be true *given that at least j-1 hypotheses are false*. Shaffer (1986) gives recursive methods for finding the values t_j for several types of testing situations. Tables of values of t_j for pairwise mean tests for m = 3,...,10 may be found in Holland & Copenhaver (1987) and Westfall & Young (1993). The methods discussed above can be modified to increase power when the hypotheses are logically related.

MODIFIED METHODS As is clear from the proof that it maintains FWE control, the Holm procedure can be modified as follows: At stage j, instead of rejecting H_j only if $p_j \leq \alpha/(n-j+1)$, H_j can be rejected if $p_j \leq \alpha/t_j$. Thus, when the hypotheses of primary interest are logically related, as in the example above, the modified sequentially-rejective Bonferroni method is more powerful than the unmodified method. For some simple applications, see Levin, Serlin, & Seaman (1994).

Hochberg & Rom (1994) and Hommel (1988) describe modifications of their Simes-based procedures for logically-related hypotheses. Of the two modifications described by Hochberg & Rom, the simpler is: proceed from i=n, n-1, n-2, etc. until

for the first time $p_i \leq \alpha/(n-i+1)$. Then reject all H_i for which $p_i \leq \alpha/t_{i+1}$. (The Rom (1990) modification of the Hochberg procedure can be improved in a similar way.) In the Hommel modification let j be the largest integer in the set $n, t_2, \ldots, t_n$, and proceed as in the unmodified Hommel procedure.

Still further modifications at the expense of greater complexity can be achieved, since it can also be shown (Shaffer 1986) that for FWE control it is necessary to consider only the number of hypotheses that can be true *given that the specific hypotheses that have been rejected are false.* Rasmussen (1993) gives algorithms that simplify application of this modification for comparing treatment means when the sample sizes are equal. Hommel (1986), Conforti & Hochberg (1987), Rom & Holland (1994), and Hochberg & Rom (1994) consider more general procedures incorporating these modifications.

COMPARISON OF PROCEDURES Among the unmodified procedures, Hommel's and Rom's are more powerful than Hochberg's, which is more powerful than Holm's; the latter two, however, are the easiest to apply (Hommel 1988,1989, Hochberg 1988, Hochberg & Rom 1994). Simulation results using the unmodified methods suggest that the differences are usually small (Holland, 1991). Comparisons among the modified procedures are more complex; see Hochberg and Rom (1994).

A CAUTION Note that all methods based on Simes' results rest on the assumption that the equality he proved for independent tests results in a conservative multiple comparison procedure for dependent tests. Thus, the use of these methods in atypical multiple test situations should be backed up by simulation or further theoretical results; for some such results, see Hochberg & Rom (1994).

Methods Controlling the FDR

The ordered p-value methods described above all are intended to provide strong control

of the FWE. Control of the FDR is a less conservative type of error control. When the test statistics are independent, Benjamini & Hochberg (1994a) show that the following step-up procedure controls the FDR at level α: If $p_j \leq j\alpha/n$, reject all H_i for $i \leq j$. (Note that the Simes equality implies that this method also controls the FWE at level α in the independent case when all null hypotheses are true, i.e. it weakly controls the FWE at level α. However, as shown by Hommel 1988, it does not control the FWE under all configurations of true and false hypotheses at level α.) A recent simulation study (Benjamini, Hochberg, & Kling 1994) suggests that the FDR is controlled at level α for the dependent tests involved in pairwise comparisons as well as for independent tests.

To see that the Benjamini-Hochberg (1994a) method can lead to many more rejections than the Hochberg step-up procedure, consider, for example, n=10, and suppose $p_n > \alpha$. Then the Hochberg procedure would reject $H_2, \ldots, H_{n-1}$ if $p_{n-1} \leq 0.5\alpha$, while the Benjamini-Hochberg (1994) FDR-controlling procedure would reject $H_2, \ldots, H_{n-1}$ if $p_{n-1} \leq 0.9\alpha$.

Williams, Jones, and Tukey (1994) give a number of examples comparing the results of applying the Benjamini-Hochberg (1994a) FDR method and several FWE-controlling methods described above. In one example they assessed significance of the changes in 8th-grade Mathematics achievement in 34 states between 1990 and 1992, as measured by the National Assessment of Educational Progress Trial State Assessment mentioned previously. Only three of the 34 changes were negative, and those were very small, so all significant results indicated improvement. Using two-sided t tests and the .05 level, 15 of the changes were significant using tests controlling only the PCE, i.e. had values $p_i \leq .05$, 6 were significant using the Bonferroni procedure, 6 again using the Hochberg (1988) procedure, and 12 using the Benjamini-Hochberg FDR-controlling procedure. In most of their examples, the Hochberg modification of the Bonferroni procedure resulted in somewhat more rejections than the simple Bonferroni procedure, but

the Benjamini-Hochberg procedure led to substantially more rejections, sometimes coming close to the number of rejections based on t-tests with no correction for multiplicity.

FDR error control is much less stringent than FWE control, but may be an attractive alternative when the number of hypotheses is very large, assuming the consequences of a small proportion of errors would not be extremely deleterious. Note, however, that in order to obtain an expected proportion of false rejections, Benjamini and Hochberg have to define a value when the denominator, the number of rejections, equals zero; they define the ratio then as zero. Then the expected proportion of false rejections, *given that some rejections actually occur,* is greater than α in some situations (in fact it necessarily equals 1 when all hypotheses are true), so more investigation of the error properties of this procedure is needed.

COMPARING NORMALLY-DISTRIBUTED MEANS

The methods in this section differ from those of the last in three respects: They deal specifically with comparisons of means, they are derived assuming normally-distributed observations, and they are based on the joint distribution of all observations. In contrast, the methods considered in the previous section are completely general, both with respect to the types of hypotheses and the distributions of test statistics, and, except for some results related to independence of statistics, utilize only the individual marginal distributions of those statistics.

Contrasts among treatment means are linear functions of the form $\sum c_i \mu_i$, where $\sum c_i = 0$. The pairwise differences among means are called simple contrasts; a general contrast can be thought of as a weighted average of some subset of means minus a weighted average of another subset. The reader is presumably familiar with the most commonly-used methods for testing the hypotheses that sets of linear contrasts equal

zero with FWE control in a one-way analysis of variance layout under standard assumptions. They will be described briefly.

Assume m treatments with N observations per treatment and a total of T observations over all treatments, let $\bar{y}_i$ be the sample mean for treatment i, and let MSW be the within-treatments mean square.

If the primary hypotheses consist of all linear contrasts among treatment means, the Scheffé method controls the familywise error rate. Using the Scheffé method (Scheffé 1953), a contrast hypothesis $\sum c_i \mu_i = 0$ is rejected if $|\sum c_i \bar{y}_i| \geq \sqrt{\sum c_i^2 (MSW/N)(m-1)F_{m-1,T-m;\alpha}}$ where $F_{m-1,T-m;\alpha}$ is the α-level critical value of the F distribution with m-1 and T-m degrees of freedom.

If the primary hypotheses consist of the pairwise differences, i.e. the simple contrasts, the Tukey method (Tukey 1953) controls the familywise error rate over this set. Using this method, any simple contrast hypothesis $\delta_{ij} = 0$ is rejected if $|\bar{y}_i - \bar{y}_j| \geq \sqrt{MSW/N}\, q_{m,T-m;\alpha}$, where $q_{m,T-m;\alpha}$ is the α-level critical value of the studentized range statistic for m means and T-m error degrees of freedom.

If the primary hypotheses consist of comparisons of each of the first m-1 means with the mth mean (e.g. of m-1 treatments with a control), the Dunnett method (Dunnett 1955) controls the familywise error rate over this set. Using this method, any hypothesis $\delta_{im} = 0$ is rejected if $|\bar{y}_i - \bar{y}_m| \geq \sqrt{2MSW/N}\, d_{m-1,T-m;\alpha}$, where $d_{m-1,T-m;\alpha}$ is the α-level critical value of the appropriate distribution for this test.

Both the Tukey and Dunnett methods can be generalized to test the hypotheses that all linear contrasts among the means equal zero, so that the three procedures can be compared in power on this whole set of tests, and similarly on the lengths of the associated confidence intervals. (These generalizations don't apply to the multistage methods described next.) Among the three, the Scheffé method is most powerful for tests on very complex contrasts, the Tukey is most powerful for tests on simple contrasts among

the first m-1 means, and the Dunnett is most powerful for tests on the contrasts δ_{im}. For discussion of these extended methods and specific comparisons, see Shaffer (1977). For a more general treatment of the extension of confidence intervals for a finite set to intervals for all linear functions of the set see Richmond (1982).

All three methods can be modified to multistage methods that give more power for hypothesis testing. In the case of the Scheffé method, if the F test is significant, the FWE is preserved if m-1 is replaced by m-2 everywhere in the expression for Scheffé significance tests (Scheffé 1970).

The Tukey method can be improved by a multiple range test using significance levels described by Tukey (1953) and sometimes referred to as Tukey-Welsch-Ryan levels; see also Einot & Gabriel (1975), Lehmann & Shaffer (1979). Begun & Gabriel (1981) describe an improved but more complex multiple range procedure based on a suggestion by Peritz (1970) using closure principles, and denoted the Peritz-Begun-Gabriel method by Grechanovsky (1993). Welsch (1977) and Dunnett & Tamhane (1992) proposed step-up methods (looking first at adjacent differences) as opposed to the stepdown methods in the multiple range procedures just described. The step-up methods have some desirable properties; see Ramsey (1981), Dunnett & Tamhane (1992), Keselman & Lix (1994), but require heavy computation or special tables for application.

The Dunnett test can be treated in a sequentially-rejective fashion, where at stage j the smaller value $d_{m-j,T-m;\alpha}$ can be substituted for $d_{m-1,T-m;\alpha}$.

Since the hypotheses in a closed set may each be tested at level α by a variety of procedures, there are many other possible multistage procedures. For example, results of Ramsey (1978), Shaffer (1981), and Kunert (1990) suggest that for most configurations of means, a multiple F-test multistage procedure is more powerful than the multiple range procedures described above for testing pairwise differences,

although the opposite is true with single-stage procedures. Other approaches to comparing means based on the joint distributions of nonoverlapping ranges have been investigated by Braun & Tukey (1983), Finner (1988), and Royen (1989,1990).

The Scheffé method and its multistage version are applicable in a straightforward way when sample sizes are unequal; simply substitute N_i for N in the Scheffé formula given above, where N_i is the number of observations for treatment i. Exact solutions for the Tukey and Dunnett procedures are possible in principle, but involve evaluation of multidimensional integrals. More practical approximate methods are based on replacing MSW/N, which is half the estimated variance of $\bar{y}_i - \bar{y}_j$ in the equal-sample-size case, with $(1/2)MSW(1/N_i + 1/N_j)$, which is half its estimated variance in the unequal-sample-size case. The common value MSW/N is thus replaced by a different value for each pair of subscripts i and j. The Tukey-Kramer method (Tukey 1953, Kramer 1956) uses the single-stage Tukey studentized range procedure with these half-variance estimates substituted for MSW/N. Kramer (1956) proposed a similar multistage method; a preferred, somewhat less conservative method proposed by Duncan (1957) modifies the Tukey multiple range method to allow for the fact that a small difference may be more significant than a large difference if it is based on larger sample sizes. See Hochberg & Tamhane (1987), who discuss the implementation of the Duncan modification and show it is conservative in the unbalanced one-way layout.

For modifications of the Dunnett procedure for unequal sample sizes, see Hochberg & Tamhane (1987).

The methods must be modified when it can't be assumed that within-treatment variances are equal. If variance heterogeneity is suspected, it is important to use a separate variance estimate for each sample mean difference or other contrast. The multiple comparison procedure should be based on the set of values of each mean difference or contrast divided by the square root of its estimated variance. The distribution of

each can be approximated by a t distribution with estimated degrees of freedom (Welch 1938, Satterthwaite 1946). Tamhane (1979) and Dunnett (1980) compared a number of single-stage procedures based on these approximate t statistics; several provided satisfactory error control.

In one-way repeated measures designs (one factor within-subjects or subjects by treatments designs), the standard mixed model assumes sphericity of the treatment covariance matrix, equivalent to the assumption of equality of the variance of each difference between sample treatment means. In standard mixed models for between-subjects-within-subjects designs there is the added assumption of equality of the covariance matrices among the levels of the between-subjects factor(s). Keselman, Keselman, & Shaffer (1991) give a detailed account of the calculation of appropriate test statistics when both these assumptions are violated, and show in a simulation study that simple single-stage multiple comparison procedures based on these statistics have satisfactory properties. See also Keselman & Lix (1994), who investigated multistage procedures and who show by simulation that both a step-up method and an adaptation of the Duncan (1957) modification, for unequal sample sizes, of the Tukey multiple range procedure have good error and power properties.

OTHER ISSUES

In this section, a number of interrelated issues of importance in making a choice among methods will be reviewed briefly.

Tests vs. Confidence Intervals

The simple Bonferroni, and the basic Scheffé, Tukey, and Dunnett methods described above are single-stage methods, and all have associated simultaneous confidence interval interpretations. When a confidence interval for a difference doesn't include zero, the hypothesis that the difference is zero is rejected, but the confidence interval gives

more information by indicating the direction and something about the magnitude of the difference or, if the hypothesis isn't rejected, the power of the procedure can be gauged by the width of the interval. In contrast, the multistage or stepwise procedures have no such straightforward confidence-interval interpretations, but intervals of a more complicated kind can sometimes be constructed. The first confidence interval interpretation of a multistage procedure was given by Kim, Stefansson, & Hsu (1988), and recently Hayter & Hsu (1994) have described a general method for obtaining these intervals with an extended discussion and a number of examples. The intervals are complicated in structure, more assumptions are required for them to be valid than for conventional confidence intervals, and the interval for each parameter depends on the values of test statistics for other parameters, as might be expected. Furthermore, although as a testing method a multistage procedure might be uniformly more powerful than a single-stage procedure, the confidence intervals corresponding to the former are sometimes less informative than those corresponding to the latter. Nonetheless, these are interesting recent results, and more along this line are to be expected.

The importance of having the usual confidence intervals derived from single-stage methods must be weighed against the increased power obtainable from multistage methods.

Directional vs Nondirectional Inference

In the examples discussed above, attention has been focused primarily on simple contrasts, testing hypotheses H_0:$\delta_{ij}=0$ vs. H_A:$\delta_{ij}\neq 0$. However, in most cases, if H_0 is rejected, it is crucial to conclude either $\mu_i > \mu_j$ or $\mu_i < \mu_j$. A number of different types of testing problems arise when direction of difference is considered.

(1) Sometimes the interest is in testing one-sided hypotheses of the form $\mu_i \leq \mu_j$ vs $\mu_i > \mu_j$, e.g. if a new treatment is being tested to see whether it is better than a standard

treatment, and there is no interest in pursuing the matter further if it is inferior.

(2) In a two-sided hypothesis test, as formulated above, rejection of the hypothesis is equivalent to the decision $\mu_i \neq \mu_j$. Is it appropriate to further conclude $\mu_i > \mu_j$ if $\bar{y}_i > \bar{y}_j$ and the opposite otherwise?

(3) Sometimes there is an a priori ordering assumption $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_m$, or some subset of these means are considered ordered, and the interest is in deciding whether some of these inequalities are strict.

Each of these situations is different, and different considerations arise. An important issue in connection with (2) and (3) above is whether it makes sense to even consider the possibility that the means under two different experimental conditions are equal. Some writers contend that a priori no difference is ever zero: for a recent defense of this position see Tukey (1991,1993). Others, including this author, believe that there is no necessity to assume that every variation in conditions must have an effect. In any case, even if one believes that a mean difference of zero is impossible, an intervention can have an effect so minute that it is essentially undetectable and unimportant, in which case the null hypothesis is reasonable as a practical way of framing the question. Whatever the views on this issue, the hypotheses in (2) are not correctly specified if directional decisions are desired. One must consider, in addition to Type I and Type II errors, the probably more severe error of concluding a difference exists but making the wrong choice of direction. This has sometimes been called a Type III error and may be the major or even the only concern in (2).

For methods with corresponding simultaneous confidence intervals, inspection of the intervals yields a directional answer immediately. For many multistage methods, the situation is less clear. Shaffer (1980) showed that an additional decision on direction in (2) does not control the FWE of Type III for all test statistic distributions, although the error is controlled for independent normally-distributed test statistics using

Holm's sequentially- rejective method. Recent simulation results (Hochberg & Parmat 1994) indicate that directional error control in the case of independent normal statistics also holds for the methods of Hochberg (1988), Hommel (1988), and Rom (1990) based on the Simes inequality. Hochberg & Tamhane (1987) describe the Shaffer (1980) results and others due to Holm (1979b); for more recent results see Finner (1990), and for a general review see Hochberg & Parmat (1994). Less powerful methods which do, however, have guaranteed Type I and/or III FWE control have been developed by Spjøtvoll (1972), Holm (1979a, improved and extended by Bauer, Hackl, Hommel, & Sonnemann 1986), Bohrer (1979), Bofinger (1985), and Hochberg (1987).

Some writers have considered methods for testing one-sided hypotheses of the type (3) (e.g. Marcus, Peritz, & Gabriel 1976, Spjøtvoll 1977, Berenson 1982.) Budde & Bauer (1989) compare a number of such procedures both theoretically and via simulation.

In another type of one-sided situation, Hsu (1981,1984) introduced a method that can be used to test the set of primary hypotheses of the form: H_i:μ_i is the largest mean. The tests are closely related to a one-sided version of the Dunnett method described above. They also relate the multiple testing literature to the ranking and selection literature.

Robustness

This is a necessarily brief look at robustness of methods based on the homogeneity of variance and normality assumptions of standard analysis of variance. Chapter 10 of Scheffé (1959) is a good source for basic theoretical results concerning these violations.

As Tukey (1993) has pointed out, an amount of variance heterogeneity that affects an overall F test only slightly becomes a more serious concern when multiple comparison methods are used, since the variance of a particular comparison may be badly

biased by use of a common estimated value. Hochberg & Tamhane (1987) discuss the effects of variance heterogeneity on the error properties of tests based on the assumption of homogeneity. Some more appropriate alternative analyses, given variance heterogeneity, are discussed above.

With respect to nonnormality, asymptotic theory ensures that with sufficiently large samples, results on Type I error and power in comparisons of means based on normally-distributed observations are approximately valid under a wide variety of nonnormal distributions. (Results assuming normally-distributed observations often are not even approximately valid under nonnormality, however, for inference on variances, covariances, and correlations.) This still leaves the issue: How large is large? In addition, alternative methods are more powerful than normal-theory-based methods under many nonnormal distributions.

A consideration of robust multiple comparison methods is beyond the scope of this article; only a couple of points will be noted.

(a) The ordered p-value methods require only accurate p-values for the individual hypotheses, so they are sometimes easier to apply than more global methods based on joint distributions of the test statistics. Note however that, if normal or other large-sample approximations are invoked, samples must be larger when more hypotheses are tested, since more extreme tails of the distributions of the relevant individual statistics are involved, and approximations are usually relatively poorer in more extreme tails.

(b) One of the major causes of poor performance of normal-theory methods is the presence of outliers, due either to errors or to heavy-tailed distributions for which extreme values are more likely than with normal distributions. Methods involving trimmed distributions and rank-based methods, among others, are often more robust in these cases. If distributions are asymmetric, estimates of the means of trimmed distributions are not estimates of the means of the original distributions; this may or may not be an issue.

Rank-based methods have discrete jumps in possible significance probabilities; with small samples and a large number of hypotheses, it may be impossible to achieve the small significance probabilities needed in applying the usual Bonferroni procedures. (See Rom 1992 for an approach to this latter problem.)

Hochberg & Tamhane (1987, Ch. 9) discuss distribution-free and robust procedures and give references to the many studies both of the robustness of normal-theory-based methods and of possible alternative methods for multiple comparisons. In addition, Westfall and Young (1993) give detailed guidance for using resampling methods to obtain appropriate error control.

Others

FREQUENTIST METHODS, BAYESIAN METHODS, AND META-ANALYSIS Frequentist methods control error without any assumptions about possible alternative values of parameters except for those that may be implied logically. Meta-analysis in its simplest form assumes that all hypotheses refer to the same parameter and it combines results into a single statement. Bayes and Empirical Bayes procedures might be thought of as intermediate in that they assume some connection among parameters and base error control on that assumption. A major contributor to the Bayesian methods is Duncan (see e.g. Duncan, 1961, 1965, Duncan & Dixon, 1983). Hochberg & Tamhane (1987) describe Bayesian approaches; see also Berry (1988). Westfall & Young (1993) discuss the relations among these three approaches.

DECISION-THEORETIC OPTIMALITY Lehmann (1957 a,b), Bohrer (1979) and Spjøtvoll (1972) defined optimal multiple comparison methods based on frequentist decision-theoretic principles, and Duncan (1961, 1965) and coworkers developed optimal procedures from the Bayesian decision-theoretic point of view. Hochberg & Tamhane (1987) discuss these and other results.

RANKING AND SELECTION This is a large related literature that is beyond the scope of this review. The methods of Dunnett (1955) and Hsu (1981, 1984), discussed above, form a bridge between the selection and multiple testing literature, and are discussed in relation to that literature in Hochberg & Tamhane (1987). Another method incorporating aspects of both approaches is described in Bechhofer, Dunnett, & Tamhane (1989).

GRAPHS AND DIAGRAMS As with all statistical results, the results of multiple comparison procedures are often most clearly and comprehensively conveyed through graphs and diagrams, especially when a large number of tests is involved. Hochberg & Tamhane (1987) discuss a number of procedures. Duncan (1955) has some illuminating geometric diagrams of acceptance regions, as do Tukey (1953), and Bohrer and Schervish (1980). Tukey (1953,1991) gives a number of graphical methods for describing differences among means; see also Hochberg, Weiss, & Hart (1982), Gabriel & Gheva (1982), and Hsu & Peruggia (1994). Tukey (1993) suggests graphical methods for displaying interactions. Schweder & Spjøtvoll (1982) illustrate a graphical method for plotting large numbers of ordered p-values that can be used to help decide on the number of true hypotheses; this approach is used by Benjamini & Hochberg (1994b) to develop a more powerful FDR-controlling method. See Hochberg & Tamhane (1987) for further references.

HIGHER-ORDER BONFERRONI AND OTHER INEQUALITIES One way of utilizing partial knowledge of joint distributions is to consider higher-order Bonferroni inequalities in testing some of the intersection hypotheses, thus potentially increasing the power of FWE-controlling multiple comparison methods. The Bonferroni inequalities are derived from a general expression for the probability of the union of a number of events. The simple Bonferroni methods using individual p-values are based on the upper bound given by the first-order inequality. Second-order approximations use joint distributions of pairs of test statistics, third-order approximations use joint distributions

of triples of test statistics, etc., thus forming a bridge between methods requiring only univariate distributions and those requiring the full multivariate distribution. See Hochberg & Tamhane (1987) for further references to methods based on second-order approximations; see also Bauer & Hackl (1985). Hoover (1990) gives results using third-order or higher approximations, and Glaz (1993) includes an extensive discussion of these inequalities. (See also Naiman & Wynn 1992, Hoppe 1993a, and Seneta 1993.) Some approaches are based on the distribution of combinations of p-values; see Cameron & Eagleson (1985), Buckley & Eagleson (1986), Maurer & Mellein (1988), and Rom & Connell (1994). A number of other types of inequalities are useful in obtaining improved approximate methods; see Hochberg & Tamhane (1987, Appendix 2).

WEIGHTS In the description of the simple Bonferroni method it was noted that each hypothesis H_i can be tested at any level α_i with the FWE controlled at $\alpha=\sum\alpha_i$. In the great majority of applications, the α_i are taken to be equal, but there may be reasons to prefer unequal allocation of error protection. For methods controlling FWE see Holm (1979a), Rosenthal & Rubin (1983), DeCani (1984), and Hochberg & Liberman (1994). Benjamini & Hochberg (1994c) extend the FDR method to allow for unequal weights, and discuss various purposes for differential weighting and alternative methods of achieving it.

OTHER AREAS OF APPLICATION Hypotheses specifying values of linear combinations of independent normal means other than contrasts can be tested jointly using the distribution of either the maximum modulus or the augmented range: see Scheffé (1959) for details. Hochberg & Tamhane (1987) discuss methods in analysis of covariance, methods for categorical data, methods for comparing variances, and experimental design issues in various areas. For further work in experimental design for multiple comparisons, see Ture (1994). Cameron & Eagleson (1985) and Buckley & Eagleson

(1986) consider multiple tests for significance of correlations. Gabriel (1968) and Morrison (1990) deal with methods for multivariate multiple comparisons. Westfall & Young (1993, Ch. 4) discuss resampling methods in a variety of situations. The large literature on model selection in regression includes many papers focusing on the multiple testing aspects of this area.

CONCLUSION

The problem of multiplicity is gaining increasing recognition, and research in the area is proliferating. The major challenge is to devise methods that incorporate some kind of overall control of Type I error while retaining reasonable power for tests of the individual hypotheses. This review, while sketching a number of issues and approaches, has emphasized recent research on relatively simple and general multistage testing methods that are providing progress in this direction.

REFERENCES

Ahmed SW. 1991. Issues arising in the application of Bonferroni procedures in federal surveys. *1991 ASA Proceedings of the Survey Research Methods Section,* pp. 344-49

Armitage, P. 1993. Interim analyses in clinical trials. See Hoppe 1993, pp. 391-402

Bauer P, Hackl P. 1985. The application of Hunter's inequality to simultaneous testing. *Biometrical J.* 27:25-38

Bauer P, Hackl P, Hommel G, Sonnemann E. 1986. Multiple testing of pairs of one-sided hypotheses. *Metrika* 33:121-27

Bauer P, Hommel G, Sonnemann E, eds. 1988. *Multiple Hypothesenprüfung. (Multiple Hypotheses Testing.)* Berlin: Springer-Verlag

Bechhofer RE. 1952. The probability of a correct ranking. *Ann. Math. Stat.* 23:139-40

Bechhofer RE, Dunnett CW, Tamhane AC. 1989. Two-stage procedures for comparing treatments with a control: elimination at the first stage and estimation at the second stage. *Biometrical J.* 31:545-61

Begun J, Gabriel KR. 1981. Closure of the Newman-Keuls multiple comparison procedure. *J. Amer. Stat. Assoc.* 76:241-45

Benjamini Y, Hochberg Y. 1994a. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Royal Stat. Soc. Ser. B,* in press

Benjamini Y, Hochberg Y. 1994b. The adaptive control of the false discovery rate. Paper submitted for publication

Benjamini Y, Hochberg Y. 1994c. Multiple hypothesis testing with weights. Paper submitted for publication

Benjamini Y, Hochberg Y, Kling Y. 1994. Controlling the false discovery rate in pairwise comparisons. Paper in preparation

Berenson ML. 1982. A comparison of several k sample tests for ordered alternatives in completely randomized designs. *Psychometrika* 47: 265-80 (Corr. 535-39)

Berry DA. 1988. Multiple comparisons, multiple tests, and data dredging: a Bayesian perspective (with discussion). In *Bayesian Statistics,* ed. JM Bernardo, MH DeGroot, DV Lindley, AFM Smith, 3:79-94. London: Oxford University Press

Bofinger E. 1985. Multiple comparisons and Type III errors. *J. Amer. Stat. Assoc.* 80:433-37

Bohrer R. 1979. Multiple three-decision rules for parametric signs. *J. Amer. Stat. Assoc.* 74:432-37

Bohrer R, Schervish MJ. 1980. An optimal multiple decision rule for signs of parameters. *Proc. Natl. Acad. Sci. USA* 77:52-56

Booth JG. 1994. Review of "Resampling Based Multiple Testing." *J. Amer. Stat. Assoc.* 89:354-55

Braun HI, ed. 1994. *The Collected Works of John W. Tukey Vol. VIII- Multiple Comparisons:1948-1983.* New York: Chapman & Hall

Braun HI, Tukey JW. 1983. Multiple comparisons through orderly partitions: the maximum subrange procedure. In *Principals of Modern Psychological Measurement: A Festschrift for Frederic M. Lord,* ed. H Wainer, S Messick, pp.55-65. Hillsdale, NJ: Erlbaum

Buckley MJ, Eagleson GK. 1986. Assessing large sets of rank correlations. *Biometrika* 73:151-57

Budde M, Bauer P. 1989. Multiple test procedures in clinical dose finding studies. *J. Amer. Stat. Assoc.* 84:792-96

Cameron MA, Eagleson GK. 1985. A new procedure for assessing large sets of correlations. *Austral. J. Stat.* 27:84-95

Chaubey YP. 1993. Review of "Resampling Based Multiple Testing." *Technometrics* 35:450-51

Conforti M, Hochberg Y. 1987. Sequentially rejective pairwise testing procedures. *J. Stat. Planning and Inference* 17:193-208

Cournot AA. 1843. *Exposition de la Théorie des Chances et des Probabilités* . Paris: Hachette. (Reprinted 1984 as vol. 1 of Cournot's *Oevres Complètes*, ed. Bernard Bru. Paris: J. Vrin.)

DeCani JS. 1984. Balancing Type I risk and loss of power in ordered Bonferroni procedures. *J. Educ. Psychol.* 76:1035-37

DeMets DL. 1987. Practical aspects in data monitoring: a brief review. *Stat. in Medicine* 6:753-60

Diaconis P. 1985. Theories of data analysis: from magical thinking through classical statistics. In *Exploring Data Tables, Trends, and Shapes,* ed. DC Hoaglin, F Mosteller, JW Tukey, pp.1-36. New York:Wiley

Duncan DB. 1951. A significance test for differences between ranked treatments in an analysis of variance. *Virginia J. Sci.* 2:172-89

Duncan DB. 1955. Multiple range and multiple F tests. *Biometrics* 11:1-42

Duncan DB. 1957. Multiple range tests for correlated and heteroscedastic means. *Biometrics* 13:164-76

Duncan DB. 1961. Bayes rules for a common multiple comparisons problem and related Student-t problems. *Ann. Math. Stat. 32:*1013-33

Duncan DB. 1965. A Bayesian approach to multiple comparisons. *Technometrics* 7:171-222

Duncan DB, Dixon DO. 1983. k-ratio t tests, t intervals, and point estimates for multiple comparisons. *Encyclopedia of Statistical Sciences,* ed. S Kotz, NL Johnson, 4:403-10. New York: Wiley

Dunnett CW. 1955. A multiple comparison procedure for comparing several treatments with a control. *J. Amer. Stat. Assoc.* 50:1096-1121

Dunnett CW. 1980. Pairwise multiple comparisons in the unequal variance case. *J. Amer. Stat. Assoc.* 75:796-800

Dunnett CW, Tamhane AC. 1992. A step-up multiple test procedure. *J. Amer. Stat. Assoc.* 87:162-70

Einot I, Gabriel KR. 1975. A study of the powers of several methods in multiple comparisons. *J. Amer. Stat. Assoc.* 70:574-83

Finner H. 1988. Abgeschlossene spannweitentests (Closed multiple range tests). See Bauer et al 1988, pp. 10-32

Finner H. 1990. On the modified S-method and directional errors. *Commun. Stat. Part A: Theory Methods* 19:41-53

Fligner MA. 1984. A note on two-sided distribution-free treatment versus control multiple comparisons. *J. Amer. Stat. Assoc.* 79: 208-11

Gabriel KR. 1968. Simultaneous test procedures in multivariate analysis of variance. *Biometrika* 55:489-504

Gabriel KR. 1969. Simultaneous test procedures--some theory of multiple comparisons. *Ann. Math. Stat.* 40:224-50

Gabriel KR. 1978. Comment on the paper by Ramsey. *J. Amer. Stat. Assoc.* 73:485-87

Gabriel KR, Gheva 1982. Some new simultaneous confidence intervals in MANOVA and their geometric representation and graphical display. In *Experimental Design, Statistical Models, and Genetic Statistics,* ed. K Hinkelmann, pp. 239-275. New York: Dekker

Gaffan EA. 1992. Review of "Multiple Comparisons for Researchers." *Brit. J. Math. Stat. Psychol.* 45:334-35

Glaz J. 1993. Approximate simultaneous confidence intervals. See Hoppe 1993b, pp. 149-166

Geller NL, Pocock SJ. 1987. Interim analyses in randomized clinical trials: ramifications and guidelines for practitioners. *Biometrics* 43:213-23

Grechanovsky E. 1993. *Comparing stepdown multiple comparison procedures.* Presented at Annu. Jt. Stat. Meet., 153rd, San Francisco

Harter HL. 1980. Early history of multiple comparison tests. In *Handbook of Statistics,* ed. PR Krishnaiah, 1:617-22. Amsterdam: North-Holland

Hartley HO. 1955. Some recent developments in analysis of variance. *Commun. Pure Appl. Math.* 8:47-72

Hayter AJ, Hsu JC. 1994. On the relationship between stepwise decision procedures and confidence sets. *J. Amer. Stat. Assoc.* 89:128-36

Heyse JF, Rom D. 1988. Adjusting for multiplicity of statistical tests in the analysis of carcinogenicity. *Biometrical J.* 30:883-96

Hochberg Y. 1987. Multiple classification rules for signs of parameters. *J. Stat. Planning and Inference* 15:177-88

Hochberg Y. 1988. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 75:800-3

Hochberg Y, Blair C. 1994. Improved Bonferroni procedures for testing overall and pairwise homogeneity hypotheses. Paper submitted for publication

Hochberg Y, Liberman U. 1994. An extended Simes test. *Stat. Prob. Letters,* in press

Hochberg Y, Parmat Y. 1994. On the problem of directional decisions. Paper submitted for publication

Hochberg Y, Rom D. 1994. Extensions of multiple testing procedures based on Simes' test. *J. Stat. Planning and Inference* , in press

Hochberg Y, Tamhane AC. 1987. *Multiple Comparison Procedures*. New York: Wiley

Hochberg Y, Weiss G, Hart S. 1982. On graphical procedures for multiple comparisons. *J. Amer. Stat. Assoc.* 77:767-72

Holland B. 1991. On the application of three modified Bonferroni procedures to pairwise multiple comparisons in balanced repeated measures designs. *Comp. Stat. Quarterly*. 6:219-31. (Corr. 7:223)

Holland BS, Copenhaver MD. 1987. An improved sequentially rejective Bonferroni test procedure. *Biometrics* 43:417-23. (Corr:43:737)

Holland BS, Copenhaver MD. 1988. Improved Bonferroni-type multiple testing procedures. *Psychol. Bull.* 104:145-49

Holm S. 1979a. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* 6:65-70

Holm S. 1979b. A stagewise directional test based on T statistics. Unpublished manuscript

Holm S. 1990. Review of "Multiple Hypothesis Testing." *Metrika* 37:206

Hommel G. 1983. Test of the overall hypothesis for arbitrary dependence structures. *Biometrical J.* 25:423-30

Hommel G. 1986. Multiple test procedures for arbitrary dependence structures. *Metrika* 33:321-36

Hommel G. 1988. A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika* 75:383-86

Hommel G. 1989. A comparison of two modified Bonferroni procedures. *Biometrika* 76:624-25

Hoover DR. 1990. Subset complement addition upper bounds -- An improved inclusion-exclusion method. *J. Stat. Planning and Inference* 24:195-202

Hoppe FM. 1993a. Beyond inclusion-and-exclusion: natural identities for P[exactly t events] and P[at least t events] and resulting inequalities. *International Stat. Rev.* 61:435-46

Hoppe FM, ed. 1993b. *Multiple Comparisons, Selection, and Applications in Biometry*. New York:Dekker

Hsu JC. 1981. Simultaneous confidence intervals for all distances from the 'best'. *Ann. Stat.* 9:1026-34

Hsu JC. 1984. Constrained simultaneous confidence intervals for multiple comparisons with the best. *Ann. Stat.* 12:1136-44

Hsu JC. 1996. *Multiple Comparisons: Theory and Methods*. New York: Chapman & Hall, in press

Hsu JC, Peruggia M. 1994. Graphical representations of Tukey's multiple comparison method. *J. Comput. Graph. Stat.* 3:143-61

Jennison C, Turnbull BW. 1990. Interim monitoring of medical trials: a review and commentary. *Statistical Science* 5:299-317

Keselman HJ, Keselman JC, Games PA. 1991. Maximum familywise Type I error rate: the least significant difference, Newman-Keuls, and other multiple comparison procedures. *Psychol. Bull.* 110:155-61

Keselman HJ, Keselman JC, Shaffer JP. 1991. Multiple pairwise comparisons of repeated measures means under violation of multisample sphericity. *Psychol. Bull.* 110:162-70

Keselman HJ, Lix LM. 1994. Improved repeated-measures stepwise multiple comparison procedures. Accepted for publication, *J. Educ. Stat.*

Kim WC, Stefansson G, Hsu JC. 1988. On confidence sets in multiple comparisons. In *Statistical Decision Theory and Related Topics IV,* ed. SS Gupta, JO Berger, 2:89-104. NY: Academic Press

Klockars AJ, Hancock GR. 1992. Power of recent multiple comparison procedures as applied to a complete set of planned orthogonal contrasts. *Psychol. Bull.* 111:505-10

Klockars AJ, Sax G. 1986. *Multiple Comparisons.* Newbury Park,CA: Sage

Kramer CY. 1956. Extension of multiple range tests to group means with unequal numbers of replications. *Biometrics* 12:307-10

Kunert J. 1990. On the power of tests for multiple comparison of three normal means. *J. Amer. Stat. Assoc.* 85:808-12

Läuter J. 1990. Review of "Multiple Hypotheses Testing." *Comput. Stat. Quarterly* 5:333

Lehmann EL. 1957a. A theory of some multiple decision problems, I. *Ann. Math. Stat.* 28:1-25

Lehmann EL. 1957b. A theory of some multiple decision problems, II. *Ann. Math. Stat.* 28:547-72

Lehmann EL, Shaffer JP. 1979. Optimum significance levels for multistage comparison procedures. *Ann. Stat.* 7:27-45

Levin JR, Serlin RC, Seaman MA. 1994. A controlled, powerful multiple-comparison strategy for several situations. *Psychol. Bull.* 115:153-59

Littell RC. 1989. Review of "Multiple Comparison Procedures." *Technometrics* 31:261-62

Marcus R, Peritz E, Gabriel KR. 1976. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* 63:655-60

Maurer W, Mellein B. 1988. On new multiple tests based on independent p-values and the assessment of their power. See Bauer et al 1988, pp. 48-66

Miller RG. 1966. *Simultaneous Statistical Inference.* New York: Wiley

Miller RG. 1977. Developments in multiple comparisons 1966-1976. *J. Amer. Stat. Assoc.* 72:779-88

Miller RG. 1981. *Simultaneous Statistical Inference.* New York: Wiley. 2nd ed.

Morrison DF. 1990. *Multivariate Statistical Methods.* New York: McGraw-Hill. 3rd ed.

Mosteller F. 1948. A k-sample slippage test for an extreme population. *Ann. Math. Stat.* 19:58-65

Naiman DQ, Wynn HP. 1992. Inclusion-exclusion-Bonferroni identities and inequalities for discrete tube-like problems via Euler characteristics. *Ann. Stat.* 20:43-76

Nair KR. 1948. Distribution of the extreme deviate from the sample mean. *Biometrika* 35:118-44

Nowak R. 1994. Problems in clinical trials go far beyond misconduct. *Science* 264:1538-41

Paulson E. 1949. A multiple decision procedure for certain problems in the analysis of variance. *Ann. Math. Stat.* 20: 95-98

Peritz E. 1970. A note on multiple comparisons. Unpublished manuscript

Peritz E. 1989. Review of "Multiple Comparison Procedures." *J. Educ. Stat.* 14:103-6

Ramsey PH. 1978. Power differences between pairwise multiple comparisons. *J. Amer. Stat. Assoc.* 73:479-85

Ramsey PH. 1981. Power of univariate pairwise multiple comparison procedures. *Psychol. Bull.* 90:352-66

Rasmussen JL. 1993. Algorithm for Shaffer's multiple comparison tests. *Educ. Psychol. Meas.* 53:329-35

Richmond J. 1982. A general method for constructing simultaneous confidence intervals. *J. Amer. Stat. Assoc.* 77:455-60

Rom DM. 1990. A sequentially rejective test procedure based on a modified Bonferroni inequality. *Biometrika* 77:663-65

Rom DM. 1992. Strengthening some common multiple test procedures for discrete data. *Stat. in Medicine* 11:511-14

Rom DM, Connell L. 1994. A generalized family of multiple test procedures. *Commun. Stat. Part A: Theory Methods,* in press

Rom DM, Holland B. 1994. A new closed multiple testing procedure for hierarchical families of hypotheses. *J. Stat. Planning & Inference* in press

Rosenthal R, Rubin DB. 1983. Ensemble-adjusted p values. *Psychol. Bull.* 94:540-41

Roy SN, Bose RC. 1953. Simultaneous confidence interval estimation. *Ann. Math. Stat.* 24:513-36

Royen T. 1989. Generalized maximum range tests for pairwise comparisons of several populations. *Biometrical J.* 31:905-29

Royen T. 1990. A probability inequality for ranges and its application to maximum range test procedures. *Metrika* 37:145- 54

Ryan TA. 1959. Multiple comparisons in psychological research, *Psychol. Bull.* 56:26-47

Ryan TA. 1960. Significance tests for multiple comparison of proportions, variances, and other statistics. *Psychol. Bull.* 57:318-28

Satterthwaite FE. 1946. An approximate distribution of estimates of variance components. *Biometrics* 2:110-14

Scheffé H. 1953. A method for judging all contrasts in the analysis of variance. *Biometrika* 40:87-104

Scheffé H. 1959. *The Analysis of Variance*. New York: Wiley

Scheffé H. 1970. Multiple testing versus multiple estimation. Improper confidence sets. Estimation of directions and ratios. *Ann. Math. Stat.* 41:1-19

Schweder T, Spjøtvoll E. 1982. Plots of P-values to evaluate many tests simultaneously. *Biometrika* 69:493-502

Seeger P. 1968. A note on a method for the analysis of significances en masse. *Technometrics* 10:586-93

Seneta E. 1993. Probability inequalities and Dunnett's test. See Hoppe 1993b, pp. 29-45

Shafer G, Olkin I. 1983. Adjusting p values to account for selection over dichotomies. *J. Amer. Stat. Assoc.* 78:674-78

Shaffer JP. 1977. Multiple comparisons emphasizing selected contrasts: an extension and generalization of Dunnett's procedure. *Biometrics* 33: 293-303

Shaffer JP. 1980. Control of directional errors with stagewise multiple test procedures. *Ann. Stat.* 8:1342-48

Shaffer JP. 1981. Complexity: an interpretability criterion for multiple comparisons. *J. Amer. Stat. Assoc.* 76:395-401

Shaffer JP. 1986. Modified sequentially rejective multiple test procedures. *J. Amer. Stat. Assoc.* 81:826-31

Shaffer JP. 1988. Simultaneous testing. In *Encyclopedia of Statistical Sciences,* ed. S Kotz, NL Johnson, 8:484-90. New York: Wiley

Shaffer JP. 1991. Probability of directional errors with disordinal (qualitative) interaction. *Psychometrika* 56:29-38

Simes RJ. 1986. An improved Bonferroni procedure for multiple tests of significance. *Biometrika* 73:751-54

Sorić B. 1989. Statistical "discoveries" and effect-size estimation. *J. Amer. Stat. Assoc.* 84:608-10

Spjøtvoll E. 1972. On the optimality of some multiple comparison procedures. *Ann. Math. Stat.* 43:398-411

Spjøtvoll E. 1977. Ordering ordered parameters. *Biometrika* 64:327-34

Stigler SM. 1986. *The History of Statistics*. Cambridge: Harvard Univ. Press

Tamhane AC. 1979. A comparison of procedures for multiple comparisons of means with unequal variances. *J. Amer. Stat. Assoc.* 74:471-80

Tatsuoka MM. 1992. Review of "Multiple Comparisons for Researchers." *Contemp. Psychol.* 37:775-76

Toothaker LE. 1991. *Multiple Comparisons for Researchers*. Newbury Park CA: Sage

Toothaker LE. 1993. *Multiple Comparison Procedures*. Newbury Park CA: Sage

Tukey JW. 1949. Comparing individual means in the analysis of variance. *Biometrics* 5:99-114

Tukey JW. 1952. Reminder sheets for "Multiple Comparisons." See Braun 1994:341-45

Tukey JW. 1953. The problem of multiple comparisons. See Braun 1994: 1-300

Tukey JW. 1991. The philosophy of multiple comparisons. *Statistical Science* 6:100-16

Tukey JW. 1993. Where should multiple comparisons go next? See Hoppe 1993b, pp. 187-207

Ture TE. 1994. Optimal row-column designs for multiple comparisons with a control: a complete catalog. *Technometrics* 36:292-99

Welch BL. 1938. The significance of the difference between two means when the population variances are unequal. *Biometrika* 25: 350-62

Welsch RE. 1977. Stepwise multiple comparison procedures. *J. Amer. Stat. Assoc.* 72:566-75

Westfall PH, Young SS. 1993. *Resampling-based Multiple Testing*. New York:Wiley

Williams GW. 1984. Time-space clustering of disease. In *Statistical Methods for Cancer Studies*, ed. RC Cornell, pp. 167-227 New York: Dekker

Williams VSL, Jones LV, Tukey JW. 1994. Controlling error in multiple comparisons, with special attention to the Trial State Assessment of Educational Progress. Paper in preparation

Wright SP. 1992. Adjusted p-values for simultaneous inference. *Biometrics* 48:1005-13

Ziegel ER. 1994. Review of "Multiple Comparisons, Selection, and Applications in Biometry." *Technometrics* 36:230-31