Annual Reviews
www.annualreviews.org/aronline

Annu. Rev. Psychol. 1995. 46:561–84
Copyright © 1995 by Annual Reviews Inc. All rights reserved

MULTIPLE HYPOTHESIS TESTING

Juliet Popper Shaffer

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

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

CONTENTS

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. This review concentrates on testing and deals with the special problems arising from the multiple aspect. The term "multiple comparisons" has come to be used synonymously with

0066-4308/95/0201-0561$05.00

Annual Reviews
www.annualreviews.org/aronline

"simultaneous inference," even when the inferences do not deal with comparisons. It is 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, thus 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 methods have been proposed for dealing with this problem, but no one solution will be acceptable for all situations. Three examples are given below to illustrate different types of multiple testing problems.

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, as 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...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; see also Shafer & Olkin 1983, Nowak 1994.)

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 a way that minimizes 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.

FACTORIAL DESIGNS The standard textbook presentation of multiple comparison issues is in the context of a one-factor investigation, where there is evidence

Annual Reviews
www.annualreviews.org/aronline

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 contrast 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 $2 \times 2 \times 2$ 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 separately at each level of other factors, thus bringing in another layer of multiplicity (see Shaffer 1991).

As noted above, Cournot clearly recognized the problems involved in multiple inference, but he considered them insoluble. Although there were a few isolated earlier relevant publications, sustained statistical attacks on the problems did not begin until the late 1940s. Mosteller (1948) and Nair (1948) dealt with extreme value problems; Tukey (1949) presented a more comprehensive approach. 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 & Bose (1953) developed another simultaneous confidence interval approach. Also in 1953, 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). Later, Lehmann (1957a,b) developed a decision-theoretic approach, and Duncan (1961) developed a Bayesian decision-theoretic approach shortly afterward. For additional historical material, see Tukey (1953), Harter (1980), Miller (1981), Hochberg & Tamhane (1987), and Shaffer (1988).

The first published book on multiple inference was Miller (1966), which was reissued in 1981, with the addition of a review article (Miller 1977). Except in the ranking and selection area, there were no other book-length treatments until 1986, when a series of book-length publications began to appear: 1. *Multiple Comparisons* (Klockars & Sax 1986); 2. *Multiple Comparison Procedures* (Hochberg & Tamhane 1987; for reviews, see Littell 1989, Peritz 1989); 3. *Multiple Hypothesenprüfung* (*Multiple Hypotheses Testing*) (Bauer et el 1988; for reviews, see Läuter 1990, Holm 1990); 4. *Multiple Comparisons for Researchers* (Toothaker 1991; for reviews, see Gaffan 1992, Tatsuoka 1992) and *Multiple Comparison Procedures* (Toothaker 1993); 5. *Multiple Comparisons, Selection, and Applications in Biometry* (Hoppe 1993b; for a review, see Ziegel 1994); 6. *Resampling-based Multiple Testing*

Annual Reviews
www.annualreviews.org/aronline

(Westfall & Young 1993; for reviews, see Chaubey 1993, Booth 1994); 7. *The Collected Works of John W. Tukey, Volume VII: Multiple Comparisons: 1948–1983* (Braun 1994); and 8. *Multiple Comparisons: Theory and Methods* (Hsu 1996).

This review emphasizes conceptual issues and general approaches. In particular, two types of methods are discussed in detail: (*a*) methods based on ordered p-values and (*b*) comparisons among normally distributed means. The literature cited offers many examples of the application of techniques discussed here.

ORGANIZING CONCEPTS

Primary Hypotheses, Closure, Hierarchical Sets, and 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. The following examples use a primary set based on differences $\mu_1, \mu_2, \ldots, \mu_m$ among the means 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 are of the form $H_{ijk\ldots}: \delta_{ijk\ldots} = 0$, indicating that all subscripted means are equal; e.g. H_{1234} is the hypothesis $\mu_1 = \mu_2 = \mu_3 = \mu_4$. 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 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}: $\mu_i = \mu_j = \mu_k$. The 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, H_{jk} because its intersection with either H_{ij} or H_{ik} also gives H_{ijk}. Note that the truth of a hypothesis implies the truth of all its proper components.

Annual Reviews
www.annualreviews.org/aronline

Any set of hypotheses in which some are proper components of others will be called a *hierarchical set.* (That term is sometimes used in a more limited way, but this definition is 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 $H_{12...m}$, or $\mu_1 = \mu_2 = ... = \mu_m$. The set of hypotheses that have no proper components represent 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 all hypotheses H_{ij} for all $i \neq j = 1,...m$, these pairwise equality hypotheses are the 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. Implicitly, attention was directed to relatively small and limited experiments. As a dramatic contrast, consider the example of large surveys and observational studies described above. 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 (see above example), in which 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). Miller (1981), Diaconis (1985), Hochberg & Tamhane (1987), and others discuss the issues involved in deciding on a family. Westfall & Young (1993) give explicit advice on methods for approaching complex experimental studies.

Because 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 (see Ahmed 1991). In a recent national report, each of the 780 pairwise differences among the 40 jurisdictions involved (states, territories, and the 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

Annual Reviews
www.annualreviews.org/aronline

report. See Ahmed (1991) for a discussion of this example and other issues in the context of large surveys.

Type I Error Control

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 (see Fligner 1984, Keselman et al 1991a). 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. This review concentrates 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 the *error rate per hypothesis, the error rate per family,* and *the error rate familywise* or *familywise error rate.*

The *error rate per hypothesis* (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. This error rate does not apply if the family size is infinite. The *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,* is described below. To make the three definitions above clearer, consider what they imply in a simple example in which each of n hypotheses $H_1,\ldots, H_n$ 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 not of the

Annual Reviews
www.annualreviews.org/aronline

α_i alone, but involves the joint distribution of the test statistics; it is smaller than or equal to the PFE, and larger than or equal to the largest α_i.

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. To see why this is not so, consider 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 by Seeger (1968), who credits the initial investigation to unpublished papers of Eklund. Sorić (1989) describes a different approach to this ratio. These papers (Seeger, Eklund, and Sorić) advocated informal consideration of the ratio; the following new approach is more formal. The *false discovery rate* (FDR) is the expected value of Q = (number of false significances/number of significances) (Benjamini & Hochberg 1994).

Power

As shown above, the error rate can be generalized in different ways when moving from single to multiple hypothesis testing. The same is true of power. Three definitions of power 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 is much greater for all-pairs than for any-pair or per-pair power (see also Gabriel 1978, Hochberg & Tamhane 1987).

P-Values and Adjusted P-Values

In testing a single hypothesis, investigators have moved away from simply accepting or rejecting the hypothesis, giving instead 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. A concept that allows generalization from the test of a single hypothesis

Annual Reviews
www.annualreviews.org/aronline

to the multiple context is the *adjusted p-value* (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. 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. These values are 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 et al (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 level α. For example, given four means, with the six hypotheses H_{ij}, $i \neq j = 1,\ldots, 4$ as the minimal hypotheses, the highest hypothesis in the hierarchy is H_{1234}, and no hypothesis below H_{1234} can be rejected unless it is rejected at level α. Assuming it is rejected, the hypothesis H_{12} cannot 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 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 control the FWE involves a simple logical argument. Consider every possible true situation, each of which can be represented as an intersection of null and alternative hypotheses. Only one of these situations can be the true one, and 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 are defined in terms of a finite family of hypotheses H_i, $i = 1,\ldots, n$, consisting of minimal hypotheses only. It is as-

Annual Reviews
www.annualreviews.org/aronline

sumed 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 is further 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 $p_1 \leq p_2 \leq \ldots \leq p_n$, with arbitrary ordering in case of ties. With the exception of the subsection on 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...or A_n) is smaller than or equal to the sum of their probabilities. Letting A_i stand for the rejection of H_i, $i = 1,\ldots, n$, this inequality is the basis of the Bonferroni methods discussed in this section.

THE SIMPLE BONFERRONI METHOD This method takes the form: 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 $\leq \alpha$ and to be 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. Multistage testing procedures can partially overcome this disadvantage. Some multistage modifications of the Bonferroni method are discussed below.

HOLM'S SEQUENTIALLY-REJECTIVE BONFERRONI METHOD The unweighted method is described here; for the weighted method, see Holm (1979). 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, $i < j$, and $p_j \leq \alpha/(n-j+1)$.

To prove that this method controls 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 $\leq \alpha$

Annual Reviews
www.annualreviews.org/aronline

[because 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 $k = 1,\ldots, n$ by $1 - (1 - \alpha)^{(1/k)}$, always $> \alpha/k$, although the difference is small for small values of α. These somewhat higher levels can also be used when the test statistics are *positive orthant dependent,* a class that includes the two-sided t statistics for pairwise comparisons of normally distributed means in a one-way layout. Holland & Copenhaver (1988) note this fact and give examples of other positive orthant dependent statistics.

Methods Based on the Simes Equality

Simes (1986) proved that if a set of hypotheses $H_1, H_2,\ldots, H_n$ are all true, and the associated test statistics are independent, then with probability $1 - \alpha$, $p_i > i\alpha/n$ for $i = 1,\ldots, n$, where the p_i are the ordered p-values, and 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, this appeared to be true only in rather pathological cases. Simes and others (Hommel 1988, Holland 1991, Klockars & Hancock 1992) have provided simulation results suggesting that the probability of the joint event is larger than $1 - \alpha$ for many types of dependence found in typical testing situations, including 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. Because 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 $j = 1,\ldots, n$, reject

Annual Reviews
www.annualreviews.org/aronline

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 is more difficult to understand and apply. Let j be the largest integer for which $p_{n-j+k} > k\alpha/j$ for all $k = 1, \ldots, j$. If no such j exists, reject all hypotheses; otherwise, reject all H_i with $p_i \leq \alpha/j$.

ROM'S MODIFICATION OF HOCHBERG'S PROCEDURE 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. Therefore, it is not necessary to protect against error in the event that five hypotheses are true and one is false, because this combination is impossible. Let t_j be the maximum number of hypotheses that are 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 (see also Holland & Copenhaver 1987, Westfall & Young 1993). The methods discussed above can be modified to increase power when the hypotheses are logically related; all methods in this section are intended to control the FWE at a level $\leq \alpha$.

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 et al (1994).

Hochberg & Rom (1994) and Hommel (1988) describe modifications of their Simes-based procedures for logically related hypotheses. The simpler of the two modifications the former describes is to 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

Annual Reviews
www.annualreviews.org/aronline

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. Hommel (1986), Conforti & Hochberg (1987), Rasmussen (1993), Rom & Holland (1994), and Hochberg & Rom (1994) consider more general procedures.

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 & Rom 1994).

A CAUTION All methods based on Simes's 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 (see Hochberg & Rom 1994).

Methods Controlling the False Discovery Rate

The ordered p-value methods described above provide strong control of the FWE. When the test statistics are independent, the following less conservative step-up procedure controls the FDR (Benjamini & Hochberg 1994): If $p_j \leq \alpha/n$, reject all H_i for $i \leq j$. A recent simulation study (Y Benjamini, Y Hochberg, & Y Kling, manuscript in preparation) suggests that the FDR is also controlled at this level for the dependent tests involved in pairwise comparisons. VSL Williams, LV Jones, & JW Tukey (manuscript in preparation) show in a number of real data examples that the Benjamini-Hochberg FDR-controlling procedure may result in substantially more rejections than other multiple comparison methods. However, to obtain an expected proportion of false rejections, Benjamini & Hochberg have to define a value when the denominator, i.e. the number of rejections, equals zero; they define the ratio then as zero. As a result, the expected proportion, given that some rejections actually occur, is greater than α in some situations (it necessarily equals one when all hypotheses are true), so more investigation of the error properties of this procedure is needed.

Annual Reviews
www.annualreviews.org/aronline

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, they utilize only the individual marginal distributions of those statistics.

Contrasts among treatment means are linear functions of the form $\Sigma c_i \mu_i$, where $\Sigma c_i = 0$. 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 are described briefly below.

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-treatment mean square.

If the primary hypotheses consist of all linear contrasts among treatment means, the Scheffé method (1953) controls the FWE. Using the Scheffé method, a contrast hypothesis $\Sigma c_i \mu_i = 0$ is rejected if $|\Sigma c_i \bar{y}_i| \geq \sqrt{\Sigma c_i^2 (MSW/N)(m-1) F_{m-1,T-m;\alpha}}$, 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 (1953) controls the FWE 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-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 (1955) controls the FWE 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 (for discussion of these extended methods and specific comparisons, see Shaffer 1977). Rich-

Annual Reviews
www.annualreviews.org/aronline

mond (1982) provides a more general treatment of the extension of confidence intervals for a finite set to intervals for all linear functions of the set.

All three methods can be modified to multistage methods that give more power for hypothesis testing. In 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 E Peritz [unpublished manuscript (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 step-down 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 $d_{m-1,\,T-m;\,\alpha}$.

Because 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 ranges have been investigated by Braun & Tukey (1983), Finner (1988), and Royen (1989, 1990).

The Scheffé method and its multistage version are easy to apply 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)

Annual Reviews
www.annualreviews.org/aronline

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. Hochberg & Tamhane (1987) discuss the implementation of the Duncan modification and show that 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 cannot 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 of the procedures 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. Standard models for between-subjects-within-subjects designs have the added assumption of equality of the covariance matrices among the levels of the between-subjects factor(s). Keselman et al (1991b) 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 multiple comparison procedures based on these statistics have satisfactory properties (see also Keselman & Lix 1994).

OTHER ISSUES

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 does not 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 is not 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 more complicated intervals can sometimes be constructed. The first confidence-interval interpreta-

Annual Reviews
www.annualreviews.org/aronline

tion of a multistage procedure was given by Kim et al (1988), and Hayter & Hsu (1994) have described a general method for obtaining these intervals. The intervals are complicated in structure, and more assumptions are required for them to be valid than for conventional confidence intervals. 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 results, and more along this line are to be expected.

Directional vs Nondirectional Inference

In the examples discussed above, most attention has been focused 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 $\mu_i < \mu_j$. 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 $\mu_1 \leq \mu_2 \leq \ldots \leq \mu_m$, 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 the second and third problems mentioned 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 it is not necessary 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 the second case described above 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 most important or even the only concern in the second testing situation.

Annual Reviews
www.annualreviews.org/aronline

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 the second testing situation does not control the FWE of Type III for all test statistic distributions. Hochberg & Tamhane (1987) describe these results and others found by S Holm [unpublished manuscript (1979)] (for newer results, see Finner 1990). Other less powerful methods with guaranteed Type I and/or Type III FWE control have been developed by Spjøtvoll (1972), Holm [1979; improved and extended by Bauer et al (1986)], Bohrer (1979), Bofinger (1985), and Hochberg (1987).

Some writers have considered methods for testing one-sided hypotheses of the third type discussed above (e.g. Marcus et al 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, because 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.

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 leaves the question of How large is large? In addition, alternative methods are more powerful than normal theory-based methods under many nonnormal distributions. Hochberg & Tamhane (1987, Chap. 9) discuss distribution-free and robust procedures and give references to many studies of the robustness of normal theory-based methods and of possible alternative methods for

Annual Reviews
www.annualreviews.org/aronline

multiple comparisons. In addition, Westfall & Young (1993) give detailed guidance for using robust 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 are 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 (1957a,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 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). Bechhofer et al (1989) describe another method that incorporates aspects of both approaches.

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) includes several illuminating geometric diagrams of acceptance regions, as do Tukey (1953) and Bohrer & Schervish (1980). Tukey (1953, 1991) gives a number of graphical methods for describing differences among means (see also Hochberg et al 1982, Gabriel & Gheva 1982, 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 Y Benjamini & Y Hochberg (manuscript submitted

Annual Reviews
www.annualreviews.org/aronline

for publication) to develop a more powerful FDR-controlling method. See Hochberg & Tamhane (1987) for further references.

HIGHER-ORDER BONFERRONI AND OTHER INEQUALITIES One way to use 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, Seneta 1993). Some approaches are based on the distribution of combinations of p-values (see Cameron & Eagleson 1985, Buckley & Eagleson 1986, Maurer & Mellein 1988, Rom & Connell 1994). Other types of inequalities are also 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 $\alpha = \Sigma\alpha_i$. In most applications, the α_i are equal, but there may be reasons to prefer unequal allocation of error protection. For methods controlling FWE, see Holm (1979), Rosenthal & Rubin (1983), DeCani (1984), and Hochberg & Liberman (1994). Y Benjamini & Y Hochberg (manuscript submitted for publication) 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 (for details, see Scheffé 1959). Hochberg & Tamhane (1987) discuss methods in analysis of covariance, methods for categorical data, methods for comparing variances, and experimental design issues in various areas. Cameron & Eagleson (1985) and Buckley & Eagleson (1986) consider multiple tests for significance of correlations. Gabriel (1968) and Morrison (1990) deal with methods for

Annual Reviews
www.annualreviews.org/aronline

multivariate multiple comparisons. Westfall & Young (1993, Chap. 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 field of multiple hypothesis testing is too broad to be covered entirely in a review of this length; apologies are due to many researchers whose contributions have not been acknowledged. 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.

ACKNOWLEDGMENTS

Research supported in part through the National Institute of Statistical Sciences by NSF Grant 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.

Any *Annual Review* chapter, as well as any article cited in an *Annual Review* chapter, may be purchased from the Annual Reviews Preprints and Reprints service. 1-800-347-8007; 415-259-5017; email: arpr@class.org

Literature Cited

Ahmed SW. 1991. Issues arising in the application of Bonferroni procedures in federal surveys. *1991 ASA Proc. Surv. Res. Methods Sect.*, pp. 344–49

Bauer P, Hackl P. 1985. The application of Hunter's inequality to simultaneous testing. *Biometr. 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 (In German and English)

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. *Biometr. J.* 31:545–61

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

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

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

Annual Reviews
www.annualreviews.org/aronline

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 Univ. Press

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

Bohrer R. 1979. Multiple three-decision rules for parametric signs. *J. Am. 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. Am. 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. Am. Stat. Assoc.* 84:792–96

Cameron MA, Eagleson GK. 1985. A new procedure for assessing large sets of correlations. *Aust. 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. Plan. Infer.* 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 *Oeuvres Complètes,* ed. B Bru. Paris: Vrin

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

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. *Va. 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. In *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. Am. Stat. Assoc.* 50: 1096–1121

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

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

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

Finner H. 1988. Abgeschlossene Spannweitentests (Closed multiple range tests). See Bauer et al 1988, pp. 10–32 (In German)

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. Am. 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. Am. Stat. Assoc.* 73:485–87

Gabriel KR, Gheva D. 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. 239–75. New York: Dekker

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

Glaz J. 1993. Approximate simultaneous confidence intervals. See Hoppe 1993b, pp. 149–66

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 Statis-*

Annual Reviews
www.annualreviews.org/aronline

tics, 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. Am. Stat. Assoc.* 89: 128–36
Hochberg Y. 1987. Multiple classification rules for signs of parameters. *J. Stat. Plan. Infer.* 15:177–88
Hochberg Y. 1988. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 75:800–3
Hochberg Y, Liberman U. 1994. An extended Simes test. *Stat. Prob. Lett.* In press
Hochberg Y, Rom D. 1994. Extensions of multiple testing procedures based on Simes' test. *J. Stat. Plan. Infer.* 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. Am. 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. *Comput. Stat. Q.* 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. 1979. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* 6: 65–70
Holm S. 1990. Review of "Multiple Hypothesis Testing." *Metrika* 37:206
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. Plan. Infer.* 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. *Int. 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
Keselman HJ, Keselman JC, Games PA. 1991a. 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. 1991b. 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. *J. Educ. Stat.* In press
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. New York: Academic
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. Am. Stat. Assoc.* 85:808–12
Läuter J. 1990. Review of "Multiple Hypotheses Testing." *Comput. Stat. Q.* 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

Annual Reviews
www.annualreviews.org/aronline

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. Am. 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. 1989. Review of "Multiple Comparison Procedures." *J. Educ. Stat.* 14:103–6
Ramsey PH. 1978. Power differences between pairwise multiple comparisons. *J. Am. 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. Am. 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, Connell L. 1994. A generalized family of multiple test procedures. *Commun. Stat. Part A: Theory Methods, 23.* In press
Rom DM, Holland B. 1994. A new closed multiple testing procedure for hierarchical families of hypotheses. *J. Stat. Plan. Infer.* 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. *Biometr. 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. Am. 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. Am. Stat. Assoc.* 76:395–401
Shaffer JP. 1986. Modified sequentially rejective multiple test procedures. *J. Am. 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
Soric B. 1989. Statistical "discoveries" and effect-size estimation. *J. Am. 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 proce-

Annual Reviews
www.annualreviews.org/aronline

dures for multiple comparisons of means with unequal variances. *J. Am. 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, pp. 341–45

Tukey JW. 1953. The problem of multiple comparisons. See Braun 1994, pp. 1–300

Tukey JW. 1991. The philsophy of multiple comparisons. *Stat. Sci.* 6:100–16

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

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. Am. Stat. Assoc.* 72: 566–75

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

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