Comparisons of Treatments After an Analysis of Variance in Ecology

R. W. Day; G. P. Quinn

Ecological Monographs, Vol. 59, No. 4 (Dec., 1989), 433-463.

Stable URL:
http://links.jstor.org/sici?sici=0012-9615%28198912%2959%3A4%3C433%3ACOTAAA%3E2.0.CO%3B2-H

Ecological Monographs is currently published by The Ecological Society of America.

Ecological Monographs, 59(4), 1989, pp. 433–463

COMPARISONS OF TREATMENTS AFTER AN ANALYSIS OF VARIANCE IN ECOLOGY[1]

R. W. DAY
Department of Zoology, University of Melbourne, Parkville, Victoria, 3052, Australia

AND

G. P. QUINN[2]
School of Biological Sciences, University of Sydney, New South Wales, 2006, Australia

Abstract. The statistical literature on tests to compare treatments after the analysis of variance is reviewed, and the use of these tests in ecology is examined. Monte Carlo simulations on normal and lognormal data indicate that many of the tests commonly used are inappropriate or inefficient. Particular tests are recommended for unplanned multiple comparisons on the basis of controlling experimentwise type I error rate and providing maximum power. These include tests for parametric and nonparametric cases, equal and unequal sample sizes, homogeneous and heterogeneous variances, non-independent means (repeated measures or adjusted means), and comparing treatments to a control. Formulae and a worked example are provided. The problem of violations of assumptions, especially variance heterogeneity, was investigated using simulations, and particular strategies are recommended. The advantages and use of planned comparisons in ecology are discussed, and the philosophy of hypothesis testing with unplanned multiple comparisons is considered in relation to confidence intervals and statistical estimation.

Key words: analysis of variance; assumptions; confidence intervals; homogeneity of variance; log transformation; Monte-Carlo simulations; multiple comparisons; planned comparisons; significance tests.

INTRODUCTION

Analyses of variance (ANOVAs) are very common in the recent ecological literature, and unplanned multiple comparison procedures (UMCPs) to compare treatment means often follow the ANOVA. Not only are UMCPs sometimes inappropriate, but the two UMCPs most commonly used, the Student-Newman-Keuls (SNK) and Duncan's multiple range tests, have not been recommended by most statisticians for many years (Ryan 1959, Scheffé 1959, Petrinovich and Hardyck 1969, Einot and Gabriel 1975). The criticisms are related to the type of error rate used and imply that conclusions drawn using these tests are based on inconsistent criteria. Means pronounced different at a significance level of 5% by one test may not be significantly different at this level using another test. Another problem is that the commonly used tests are not robust to violations of assumptions, and more robust tests should be used, for example, when variances may not be equal. Yet there is often no explanation in ecological papers of why a procedure was chosen, or enough information given for the readers to judge the strength of the conclusions, or interpret the results themselves. Sometimes even the method used is not properly specified.

The subject of comparisons after, or included in, the ANOVA has been a rapidly developing field in statistics, and many standard statistical texts appear to have glossed over the issue to some extent, by describing only a few common methods rather than discussing the issues involved and the alternatives available (more complete treatments are presented in Winer 1971, Sokal and Rohlf 1981, Keppel 1982, Miller 1986). Multiple comparisons have been reviewed by statisticians (e.g., O'Neill and Wetherill 1971, Thomas 1973, Miller 1981, Stoline 1981, Games et al. 1983, Hochberg and Tamhane 1987) and have been discussed in articles in many fields, e.g., agronomy (Chew 1976, Baker 1980), animal production and veterinary science (Gill 1973, Waldo 1976, Cox 1980), entomology (Jones 1984, Perry 1986), medicine (Rosen and Hoffman 1978, Salsburg 1985), plant pathology (Madden et al. 1982), and psychology (Jaccard et al. 1984, Klockers and Sax 1986). These papers have examined either the relative merits of different UMCPs or whether UMCPs should be used at all, leading to a more general discussion of the use-

[1] Manuscript received 24 August 1987; revised 29 July 1988; accepted 24 October 1988; final version received 14 December 1988.

[2] Present address: Department of Zoology, Monash University, Clayton, Victoria, 3168, Australia.

fulness of statistical hypothesis testing (e.g., Jones and Matloff 1986).

It is important for ecologists to judge whether the biological data fit the assumptions of the analysis closely enough. This depends on how robust the method used is to deviations from its assumptions. Such information is hard to find in textbooks, but there are many recent studies of this issue, and some new robust methods, both for ANOVA and comparisons of means. The assumption of equal variances is particularly important. Biological data often follow a lognormal distribution, presumably because of underlying multiplicative processes. Such processes lead to a relation between the mean and the variance, so that the samples from different treatments not only have non-normal distributions but also unequal variances. The effects of this on the ANOVA and tests to compare means has seldom been studied.

Papers written for biologists on comparing treatments have been limited in their coverage of the problem (e.g., not discussing unequal sample sizes, variance heterogeneity, or nonparametric methods) and their coverage of the literature, particularly many recent statistical papers. Nonetheless, the common theme emerging from all these papers is that different methods have been designed for different purposes, and the best method to use for a particular experiment depends on the precise objectives of the experiment. We suspect that ecologists often do not understand clearly the purpose of the method they use.

The objectives of this paper are: (a) to explain the concepts underlying the methods to compare treatments and to summarize the relevant statistical literature, much of which is very recent; (b) to determine the most commonly used methods in ecology, and the situations in which they are used; (c) to show, using Monte-Carlo simulations, what the commonly used UMCPs do and how they are affected by variance heterogeneity; (d) to demonstrate the effects of lognormal data and the need to detect this situation; and (e) to discuss which methods are most appropriate for particular situations. We intend this paper to provide practical advice on what test to use. The formulae are in Appendix 1 and a worked example is in Appendix 2. References to textbooks (where possible) and papers are given for all tests mentioned.

Relevant concepts

Statistical terms and notation will be kept to a minimum, but some basic concepts are required to discuss the merits of different procedures, and some of these appear to be poorly understood by many ecologists, despite the efforts of Underwood (1981) and various texts. The concepts in this section also apply to nonparametric comparisons of treatment medians.

Two types of error can occur when tests are used to compare treatments:

Type I: The means for two treatments are declared significantly different when, in fact, the population means for the treatments are equal.

Type II: The means for two treatments are not declared different when, in fact, the population means differ.

Type I errors are measured in two ways. Consider an experiment with four treatments with equal population means A, B, C, D, and an analysis in which the means of samples from each treatment are compared using multiple *t* tests at the 5% level. The type I error rate per comparison is the probability of a type I error in a single comparison. The probability of a type I error when A and B are compared is 5%, and the same is true when B and C are compared, etc.

The experimentwise type I error rate (EER) is defined as the probability of one or more type I errors when all of the comparisons in an experiment are made. For multiple *t* tests the EER may be roughly estimated using the binomial probability formula by treating the comparisons as independent. In the example there are six comparisons and the estimated EER is $\approx$26%. The EER can only be reduced to 5% by decreasing the error rate per comparison.

Now consider an experiment where three treatments have equal population means A, B, and C, while D is different from the other three. In comparing the means a type I error can only occur when comparing A, B, and C, so that the EER using multiple *t* tests would be $\approx$14%. The error rate per comparison does not have to be decreased as much in this case to reduce the experimentwise error rate to 5%. The stepwise UMCPs described below make use of this fact.

In a factorial experiment where means are considered in groups, one refers to the familywise error rate for each group rather than the EER. Huitema (1980) points out that if an ANOVA *F* test is carried out at the 5% level, then this type I error rate of 5% refers to the probability that one or more of all the possible comparisons between the treatments will be incorrectly declared significant, i.e., the *F* test uses an EER (or a familywise error rate in factorial experiments).

The power of a test is the probability of not making a type II error if the treatments differ (i.e., the probability of finding real differences), and this is measured on a per-comparison basis. The probability of a type II error depends on how different the treatments are in relation to the variability within treatments in the data, the sample size, and which method is used. If the type I error rate per comparison is very low then the probability of a type II error will be high, and the test will have a low power.

Heterogeneity of variance may affect both the type I and type II error rates. Box (1954) showed that the effect of heterogeneous variances on the type I error rate of the ANOVA *F* test would be approximately proportional to a coefficient of variation of the variances (Box's coefficient). For a given range of variances, Box's coefficient is largest when one variance is large and the rest are small. When sample sizes are unequal there is an additional important effect on the error rate, which depends on the ratio of the unweighted (ignoring sample sizes) to the weighted (by the degrees of freedom of each sample) mean of the variances (Box 1954). This "bias ratio" reflects the degree to which small sample sizes are paired with large variances.

Comparisons of means may be planned or unplanned. Planned comparisons are a selected subset of the possible contrasts between means or combinations of means, chosen before the experiment is done so that they are not suggested by the results. They should reflect a rational set of specific hypotheses about the treatments, which flows from the design of the experiment or the knowledge and interests of the researcher (Warren 1979). For example, in an experiment on the effects of no food and two types of food on growth rates, a pertinent hypothesis is that growth is greater in treatments with food. This is measured by the difference between the mean for no food and the average of the other two means. Another common example is where a researcher is interested in trends across levels of some factor, or simply hypothesizes that the treatment means will be ordered in some way. Ideally, planned comparisons should be orthogonal, i.e., they should test completely separate hypotheses (Sokal and Rohlf 1981) and thus provide independent pieces of information (Keppel 1982). The number of hypotheses possible in an orthogonal set is no more than the number of degrees of freedom available.

Unplanned comparisons occur when the researcher has only a general question, such as "Are there any differences?" to ask and explores the results to find what pattern of differences emerges. The common strategy here is to compare all the means in pairs, using unplanned multiple comparison procedures (UMCPs). Such comparisons are not orthogonal, and the objective is usually to see whether any of the comparisons are significant, so as to base conclusions on these results. Unplanned tests are called "protected" when they are applied only if the ANOVA F test is significant, i.e., as a two-stage procedure. The EER of protected tests is measured for the whole two-stage procedure (Carmer and Swanson 1973, Miller 1981). Some authors have suggested that this protection ensures control of the EER (see discussions in Keppel 1982, Zwick and Marascuilo 1984).

Methods for comparisons

This section will describe the available procedures for making comparisons of treatments and their rationale. As a complete, up-to-date reference source is not available elsewhere, we provide references to the relevant literature; we also provide technical information, including formulae for the tests in Appendix 1. The symbol a will be used for the chosen (i.e., nominal) significance level, e.g., .05 or 5%, m for the number of treatments or means, r for the number of comparisons and "df" for degrees of freedom.

Planned comparisons.—Planned comparisons use a per-comparison error rate. They may be pairwise comparisons, where two means are compared, or contrasts, where combinations of several means are compared, e.g., the average of two means is compared to a third. In either case they are significance tests of focused questions, as opposed to omnibus or unfocused tests (Rosenthal and Rosnow 1985).

The t or F statistics are used for parametric tests. The F method has the advantage that a sum of squares for each comparison can be put in the ANOVA table for the overall analysis, showing how the variation is partitioned between (orthogonal) comparisons. These F test comparisons are exactly equivalent to t test comparisons. Both methods assume equal variances and use the mean square in the denominator of the overall ANOVA F test to estimate the standard error of the comparison (SE$_c$). For nonparametric tests, the Mann-Whitney-Wilcoxon U statistic appears to be the simplest pairwise test to apply; and the Spjøtvoll method is best for contrasts (see Hollander and Wolfe 1973). These also require equal variances (the assumption being that the distributions only differ in location).

When treatment variances are unequal, the use of a pooled estimate of variance (e.g., the denominator mean square from the ANOVA, MS$_d$) in calculating the standard error of the comparison (SE$_c$) is inappropriate. Parametric methods robust to unequal variances include the Behrens-Fisher t test (Winer 1971, Snedecor and Cochran 1980), which uses the correct standard error (SE$_c$*), and special tables in Fisher and Yates (1953). However, standard t tables can be used in approximate tests: Welch's (1938) test uses adjusted degrees of freedom (df*), which were generalized by Satterthwaite (1946) for use in complex contrasts (see Winer 1971, Keppel 1982); Welch's (1947) slightly different adjustment (Winer 1971, Keppel 1982); and Cochran's approximate t test (see Snedecor and Cochran 1980, Sokal and Rohlf 1981). A robust version of the Mann-Whitney U test has also been described recently (Fligner and Policello 1981), which requires only that the distributions under each treatment are symmetrical. We call this the Fligner-Policello test.

Parametric unplanned comparisons.—The commonly used methods rely on the Studentized Range statistic (Q), or on the F and t statistics (which are equivalent for comparisons). Other methods use the Studentized Maximum Modulus (SMM), or the Studentized Augmented Range (SAR), which are tabulated in Rohlf and Sokal (1981). However most methods can be adapted for use with any of the statistics (Einot and Gabriel 1975). Kurtz et al. (1965) noted that the statistics vary in the kinds of differences between means to which they are most sensitive: (1) for comparisons involving combinations of many means (e.g., regression comparisons) the F statistic is most powerful; (2) for simultaneous pairwise comparisons of means the Q statistic is most powerful; and (3) the SMM statistic is most powerful for detecting treatments that are different from the remaining group of treatments. Einot and Gabriel (1975) showed that the F is more powerful than the Q statistic for stepwise UMCPs such as the Student-Newman-Keuls (SNK) procedure, but the difference is very small.

1. *Equal sample sizes and variances.*—The simplest test is the LSD (Least Significant Difference) test, which uses a t value from tables with the df of the standard error, SE_c. The LSD test and multiple t tests set the type I error rate at a per-comparison level of a, and as a result they are the most powerful tests available. Fisher (1935) proposed the Protected LSD test, where the individual comparisons are tested only if the ANOVA F test is significant. Note that the test described in Winer (1971) as a modified LSD approach suggested by Fisher, is a form of the Bonferroni method below.

The Bonferroni method involves adjusting the significance level per comparison, using the Bonferroni inequality, to ensure the EER is always below the level chosen (a). This is usually applied to the t test. A more powerful Bonferroni method, the Dunn-Sidák method, is described by Ury (1976) and Sokal and Rohlf (1981). These methods are designed for comparisons involving combinations of means as well as pairwise comparisons, provided the number of comparisons to be made (r) is fixed in advance.

Scheffé's test (Scheffé 1953) is designed, like the Dunn-Sidák method, for all possible comparisons, including both pairwise comparisons and contrasts. There is a very large number of ways to compare combinations of means, whereas the number of pairwise comparisons is limited to $r = m(m - 1)/2$, where $r =$ the number of possible comparisons to be made and $m =$ the number of treatments (or means) in the experiment. Scheffé's method, therefore, adjusts the type I error rate per comparison to a very low level to keep the EER at the chosen level (a); and it lies below a when there is only a limited number of possible comparisons.

Tukey's method (Tukey 1953), variously called the honestly significant difference (HSD) test, the T method, Tukey's A, and Tukey's w method, is usually used with the Q statistic. This test is designed for comparing each pair of means. Like each of the tests above, the critical value for each comparison is the same, so that each comparison has the same chance of a type I error. Such tests are called "*simultaneous tests.*"

There are many *stepwise tests,* which are designed to compare pairs of means to find where differences occur, or to detect groups of equivalent means. The means are arranged in order, for example A, B, C, D. The difference between A and D is tested using a "4-mean significance level," which is the probability of falsely rejecting the hypothesis that all four means are equal. If this test is significant, then groups of three means (e.g., A, B, C and B, C, D) are tested, using a "3-mean significance level." This process is continued until pairs of means are tested, but no further tests are done on means which lie within a nonsignificant group. The stepwise tests differ in how the adjusted significance level is set. They include the familiar Duncan's multiple range (new) test (Duncan 1955), and the Student-Newman-Keuls (SNK) test (Newman 1939, Keuls 1952). Ryan (1960) proposed a stepwise test, using adjusted significance levels on any two-sample test such as the t or Mann-Whitney U tests. The method is not appropriate for multisample statistics such as Q or F. Einot and Gabriel (1975) described a modification of Ryan's method for the Q and F statistics which we illustrate and call "Ryan's Q (or F) test." Welsch (1977) also proposed a number of modifications of Ryan's method. One of these is Welsch's step-up, or GAPA, test (see Sokal and Rohlf 1981) where pairs of means are tested first, then groups of three means, etc. A test on a larger group is automatically significant if it contains means which have already been declared different. Ramsey (1978) proposed a further revision of Ryan's test, which he unfortunately called "Ryan's procedure." This revised Ryan's test is slightly more complex to use than the Ryan's Q test we illustrate.

The adjusted levels of significance (b) are highest for Duncan's test and lowest for Ryan's test. Following Begun and Gabriel (1981), values of $Q_b(P, \mathrm{df})$ for Ryan's Q test can be obtained by interpolation from available tables of the Q statistic (see Appendix 1).

A number of new stepwise UMCPs have been developed recently, but will not be described in detail here. These include Shaffer's (1979) modification of stepwise procedures, which replaces the comparison of means farthest apart by an ANOVA F test; Peritz's test (Einot and Gabriel 1975), which is a mixture of the SNK test and Ryan's test (Begun and Gabriel [1981] provide a computer algorithm); and a model testing procedure described by Ramsey (1981), who has described and evaluated all these tests. They are more complex to apply than those described above and in Appendix 1.

The Waller-Duncan k-ratio test (Waller and Duncan 1969) employs a different approach from the other tests. The critical value used depends on (1) a chosen ratio of the seriousness of type I and type II errors, which corresponds to the chosen significance level, and (2) the magnitude of the ANOVA F value. It is thus like an elaborate protected LSD test.

2. *Unequal sample sizes.*—Most UMCPs can be simply modified for unequal sample sizes. The usual modification of the t test is to replace n (sample size) in the formula for the standard error (SE_c) by the harmonic mean of the sample sizes in the comparison (e.g., Sokal and Rohlf 1981). This is equivalent to using the formula for complex comparisons, and therefore the same substitution is used for the Bonferroni (or Dunn-Sidák) and Scheffé methods, which were designed for

complex comparisons. Tukey (1953) proposed this substitution for Tukey's test, and Kramer (1956) proposed it for the Duncan and SNK tests. It has become known as the Tukey-Kramer (T-K) or Kramer's method (see Sokal and Rohlf 1981). It can also be used for the other stepwise tests described above.

More recent methods include the GT2 method (Hochberg 1974), Gabriel's (1978) approximation to the GT2 method, and the T′ method (Spjøtvoll and Stoline 1973). These methods are illustrated by Sokal and Rohlf (1981). Winer (1971) and Snedecor and Cochran (1980) proposed replacing *n* by the harmonic mean of all the sample sizes in Tukey's and the stepwise tests. All these methods, and some other less-used ones, are discussed by Dunnett (1980*a*) and Stoline (1981).

3. *Heterogeneous variances (equal or unequal sample sizes).*—When treatment variances are unequal, the pooled estimate of variance (MS_d) cannot be used to calculate the standard error of the comparison (SE_c). For all pairwise comparisons of means, simulations by Tamhane (1979) and Dunnett (1980*b*) have reduced the choice of suitable UCMPs to three: the GH (Games and Howell 1976, Sokal and Rohlf 1981), *T*3 (Tamhane 1977 modified by Dunnett 1980*b*) and *C* methods. The last was developed by Dunnett (1980*b*) from Cochran's approximate *t* test. For complex comparisons, formulae for $\text{SE}_c{}^*$ and df* are provided in Appendix 1 (under Planned Comparisons) and these can be used with the Scheffé, Bonferroni, or Dunn-Sidák methods.

4. *Non-independence.*—Adjusted means in an analysis of covariance (ANCOVA) are not independent. The standard UMCPs are not appropriate (Bancroft 1968, Neter and Wasserman 1974, Miller 1981) because they require that the covariances as well as the variances be homogeneous, and this will not generally be the case (Scheffé 1959). Covariance heterogeneity leads to inflated type I error rates (Renner and Ball 1983). If the assumption of equal regression coefficients (or slopes) holds, planned comparisons, Scheffé's test, and the Bonferroni methods can still be used, if the standard error of the comparison (SE_c) is increased according to the variation among means for the covariate. Sokal and Rohlf (1981) suggested using the GT2 procedure described earlier.

Other special techniques have been developed for comparing means adjusted in ANCOVA. Thigpen and Paulson (1974) developed a simultaneous test procedure, similar to Tukey's HSD test, based on a generalized Studentized Range distribution, which was extended by Bryant and Paulson (1976) to handle designs other than one-way ANCOVA. Huitema (1980) provided an excellent description of this Bryant-Paulson generalization of Tukey's test; it can be also used as a stepwise test, with the generalized Studentized Range distribution (Bryant and Paulson 1976) or with special Duncan's multiple range tables (Bryant and Bruvold 1980). Recently, Hochberg and Varon-Salomon (1984) described a Tukey-Kramer procedure, using the standard Studentized Range distribution, which is conditional on the values of the covariate; the standard error of the comparison will vary, therefore, depending on which pair of adjusted means is being compared.

Wilcox (1987) pointed out that in an ANCOVA, procedures that first test for equal slopes and then equal intercepts (or adjusted means) do not control the EER among all the hypotheses tested. He described a technique for simultaneous pairwise comparisons of slopes and intercepts, analogous to the Tukey-Kramer method, that controls the EER.

Treatment means for the repeated factor in repeated-measures ANOVA are also not independent. We know of no special tests developed for use with repeated-measures ANOVA. Winer (1971) recommended using a standard procedure such as the SNK test, but this is only applicable if the variances and covariances are truly homogeneous.

It is also important to note that in an analysis where a number of tests are carried out using the same estimate of error variation (SE_c), the tests are not independent. This problem arises both in tests of numerous fixed factors in an ANOVA and in testing numerous comparisons using methods which assume homogeneous variances. Hurlbert and Spiegel (1976) showed that the problem becomes serious when the sum of the df of the comparisons is close to the df of the error estimate. This may occur when very few replicates are used, or in factorial designs when a random factor has few levels.

Nonparametric unplanned comparisons.—Nonparametric methods are appropriate if the population distributions are not fairly close to normal (e.g., if they are multimodal; see Discussion and Recommendations: Assumptions). Standard nonparametric methods assume that the distributions under each treatment differ only in location, i.e., median. This requires the shape and variances of distributions to be the same in all treatments, either on the original scale of measurement or on some transformed scale. Most multiple comparison procedures can be applied in the nonparametric case, using treatment rank sums (or mean ranks or placement sums) instead of means, and a nonparametric statistic. There are two broad groups of nonparametric UMCPs for pairwise comparisons that use two quite different approaches. One group uses *joint rankings,* i.e., each pairwise comparison is based on the ranks for all *m* treatments in the study. The result of the comparison of each pair of treatments depends on the data from the other $m - 2$ treatments, a situation not found in any of the commonly used parametric UMCPs. These tests usually calculate the difference in mean ranks, and include the simultaneous Nemenyi test (Nemenyi 1963, cited in Hollander and Wolfe 1973), which uses either exact tables (Hollander and Wolfe 1973, Damico and Wolfe 1987) or an extension of the Kruskal-Wallis statistic in a Scheffé test. Large sample approximations of this test use the Q statistic (Miller 1981, 1986), or the unit normal z statistic (Dunn 1964, Hollander and Wolfe 1973), with significance levels adjusted according to the Bonferroni inequality.

Zar (1974) described an SNK-type stepwise version of this test, and Campbell and Skillings (1985) have described two other stepwise versions: an "*all-subset*" *procedure* (if a set of m treatments is declared significant, then all possible subsets of $m - 1$ treatments are tested without ordering them in any way) and an "*ad hoc*" *procedure* (which orders the treatments according to rank sums to determine which subsets are tested, just as for means in a parametric stepwise test). Both use Ryan's adjusted significance levels; the ad hoc procedure is much simpler than the all-subset test (Campbell and Skillings 1985). We illustrate the Nemenyi (Joint-Rank) test and the stepwise ad hoc test (which we call the "Joint-Rank Ryan test").

The other group uses *pairwise rankings*, i.e., re-ranking the data for each pair of treatments being compared. The test for each pair of treatments does not depend on the other treatments in the study, as is always the situation in parametric procedures. This group usually calculates the maximum or minimum rank sum (or placements) and uses the Wilcoxon or Mann-Whitney U statistic. It includes the simultaneous Steel-Dwass test (Steel 1960, Miller 1981), using exact tables or a large sample approximation based on the Q statistic, and the version of this test using "placements" in Sokal and Rohlf (1981). Conover (1980) described a simultaneous method using the t statistic in an LSD-type test. Campbell and Skillings (1985) described a stepwise all-subset version of the Steel-Dwass test and suggested a stepwise ad hoc procedure (see above), both using Ryan's adjusted significance levels. We illustrate the Steel-Dwass test, and a stepwise ad hoc version (called "Steel-Dwass Ryan"), which orders the treatments according to sample medians.

The exact tables for the Nemenyi Joint-Rank test, and Dunn's (1964) generalization of both the pairwise and joint rank methods with a large sample approximation, allow unequal sample sizes (see also Miller 1981, 1986).

Comparing treatments with a control.—A special group of UMCPs is designed for the case where a control is to be compared with each of the other $m - 1$ treatments. Here there are $m - 1$ comparisons rather than $m(m - 1)/2$ as in all pairwise comparisons, so that a more powerful test can be used while keeping the EER at α. Parametric tests of this kind include Dunnett's test (Dunnett 1955) described in Winer (1971) and Zar (1974); Dunnett (1964, 1985) discussed unequal sample sizes and variances, and Shaffer (1977) extended this test to include contrasts among the treatments. Williams' (1972) procedure is suitable for comparing dose levels to a zero-dose control. The nonparametric Steel's test (Steel 1959) using pairwise ranks is described in Winer (1971) and Miller (1981), and was extended by Fligner (1984) for unequal sample sizes and contrasts among the treatments. Miller (1981) described a simpler "sign test" version and alternative tests using joint ranks are described in Hollander and Wolfe (1973).

Methods

Literature search

The journals *Ecology*, *Journal of Experimental Marine Biology and Ecology*, *Marine Biology*, and *Oecologia (Berlin)* were examined, and all papers published in the years 1982 to 1984 inclusive plus a second sample from late 1986 were reviewed. Of the total of 3350 papers, 529 contained some comparison of treatments after, or instead of, an analysis of variance (or its nonparametric equivalent). The following information was recorded from these papers:

- whether the analysis was parametric or nonparametric;
- which procedure was used, and whether it was used independently of a significant ANOVA F test or nonparametric equivalent;
- whether a homogeneity-of-variance test was carried out;
- the number of treatments compared, and the sample size of each treatment;
- whether all possible pairs of treatments, or some subset of these were compared;
- whether non-independent means (e.g., in repeated measures ANOVAs, or adjusted means after ANCOVA) were involved;
- any reference to statistical texts or papers for the particular test used.

Papers in which it was difficult to determine what type of analysis was carried out (unfortunately common) were only included in the present review when it was clear that some sort of multiple comparison had been used. In 26 of the 1986 papers, sample sizes and variance estimates were provided or could be calculated from the information, and these were used to determine the values of Box's coefficient of variance heterogeneity and Box's bias ratio that are found in practice.

Papers describing or evaluating UMCPs were examined in the statistical literature and elsewhere, to determine what is known of these procedures. Miller (1981) provided a good bibliography of statistical papers, but relevant papers on the use of UMCPs are widely scattered across journals in many disciplines.

Monte Carlo simulations

Simulations were carried out in order to illustrate how the commonly used or recommended UMCPs perform, using sample sizes and numbers of treatments that are typical of analyses in ecology. We have not adjusted each procedure to hold the experimentwise error rate (EER) constant, as suggested by Einot and Gabriel (1975). While this might provide a better comparison of the inherent properties of the procedures, ecologists use UMCPs as specified in textbooks, with different EERs. Similarly, we used sample variances to estimate population variances in calculating critical values, as this is what would occur in practice.

Simulations were run on a Hewlett Packard 200-series microcomputer, programmed in BASIC. Except where noted below, each run consisted of 12 000 experiments, in which random samples were generated for each treatment, then subjected to the ANOVA *F* test, the Kruskal-Wallis nonparametric ANOVA, and various methods to compare pairs of treatments. Nominal significance levels of 5% were used for each test. In each run the program recorded, for each UMCP, the total number of comparisons with type I and with type II errors and the number of experiments with one or more type I errors. A separate tally of these numbers was kept for those experiments where the *F* value was significant. The EER was calculated as the number of experiments with one or more type I errors divided by the number of experiments. The runs were divided into blocks of 3000 experiments to determine how much the results would vary.

Numbers with a standard normal distribution were obtained from pairs of uniform pseudo-random numbers using the method of Box and Muller (1958), and transformed to obtain samples with the population mean and variance required. The results of a spectral test (Knuth 1981) of the available linear congruential random-number generator showed that the points specified by potential pairs of pseudo-random numbers were not as closely spaced as Knuth (1981) suggested is sufficient. The method of Bays and Durham (1976) was used to shuffle the pseudo-random numbers for each experiment in a large array. This method produces numbers sufficiently random for most purposes (Knuth 1981).

1. *Normal distribution, equal variances.*—Sample sizes (n) of three, five, and nine, and experiments with three to six treatments (m) were used, as these appeared typical of the literature surveyed (see Results: Procedures Used in Ecology, below). All treatments had equal population variances. Parameters used in the simulations are shown in Table 1. In the first series of runs the treatment population means were all equal. In the second series of runs the treatments were grouped in three pairs, and the difference between pairs was set at 0.25, 0.5, 1, 1.5, 2, 2.5, 3, and 4 population standard deviations (PSDs). Because in simulations with $m = 6$ treatments the nonparametric methods required far more computer time, 4000 experiments per run were used to investigate these methods. In the third series of runs the treatments were grouped in various ways, and the difference between groups was one of the above (see Table 1). The parametric UMCPs examined were the Scheffé, Dunn-Sidák, Tukey, SNK, Duncan, and Ryan *Q* tests. The nonparametric methods were the Steel-Dwass test, the Nemenyi Joint Rank test, and two stepwise methods: the Joint Rank Ryan test (the "ad hoc" test described by Campbell and Skillings [1985]), and the Steel-Dwass Ryan test (an equivalent pairwise "ad hoc" test).

TABLE 1. Parameters used in simulations (equal variances). PSD is population standard deviation.

Simulation series	No. of means	Arrangement of means*	Distance between groups (PSDs)	Sample sizes (n)
A	6	all equal	none	3, 5, 9
B	6	2 + 2 + 2	various	3, 5, 9
C	6	3 + 3	various	3, 5, 9
D	6	various	3	5
E	5	various	3	5
F	3	3 equal, 1 + 2	3	5

* This shows the number, and size, of groups of equal means. For example, in Series F there were either three equal means or three means, of which only two were equal to each other (1 + 2).

TABLE 2. Box's coefficient of variance heterogeneity and bias ratio for initial simulations with unequal variances and three treatments. Population variances and sample sizes used are shown. ND = simulation not done.

	Equal n (3, 5, 9)	Unequal n (9, 9, 5 and 5, 5, 3)	
Variances used	Box's coefficient*	Box's coefficient	Box's bias ratio†
1, 1, 1	0.0	0.0	1.0
1, 1, 2	0.35	0.33	1.11
1, 1, 3	0.57	0.57	1.19
1, 1, 4	0.71	0.75	1.25
1, 2, 9	0.89	1.01	1.43
1, 1, 10	1.06	1.29	1.43
1, 1, 16	1.18	1.50	1.50
1, 1, 34	1.29	ND	ND

* The coefficient of variation of the variances.

† (Weighted mean of the variances)/(unweighted mean of the variances).

2. *Normal distribution, unequal variances.*—Simulations were used to illustrate the effects that heterogeneity of variances would have on the ANOVA *F* test, the Kruskal-Wallis ANOVA, and the Tukey, Ryan *Q*, and nonparametric UMCPs. Box's (1954) coefficient of variance heterogeneity guided the choice of sample sizes and population variances for the first series of simulations, with three treatments, shown in Table 2. Based on these results, further simulations were done to determine the effects of unequal sample sizes for the parametric tests (using the Tukey-Kramer modification).

3. *Lognormal distribution.*—Lognormal data may arise in practice as a result of multiplicative biological processes, so that the treatment variance is proportional to the mean. To illustrate the effect of such pro-

TABLE 3. Parametric unplanned multiple comparison tests used in papers in ecology in 1982–1984 and in 1986. NR = not recorded.

Test used	Year	Total no. papers	% of total	No. after F test	No. treatments (mean ± SD)
SNK	1982–1984	146	38	132	5.0 ± 2.5
	1986	23	29	21	5.1 ± 2.7
Duncan's	1982–1984	80	21	64	5.3 ± 2.5
	1986	15	19	13	4.2 ± 1.9
Multiple t	1982–1984	44	11	6	4.8 ± 3.6
	1986	7	9	3	4.3 ± 1.4
Tukey's	1982–1984	30	8	25	5.4 ± 3.3
	1986	13	16	12	5.4 ± 3.3
LSD	1982–1984	25	6	18	5.6 ± 2.9
	1986	5	6	5	5.4 ± 2.1
Scheffé's	1982–1984	22	6	20	4.7 ± 2.0
	1986	3	4	3	3.3 ± 1.2
Bonferroni t	1982–1984	6	2	3	NR
	1986	4	5	3	NR
Other UMCPs	1982–1984	10	2	7	NR
	1986	3	4	2	NR
Unspecified	1982–1984	22	6	17	NR
	1986	7	8	5	NR

cesses, lognormal data were generated by calculating the antilog (i.e., e^x) of normally distributed data, generated as described previously. The population means of the underlying normal samples were varied, so that the variances of the lognormal samples varied in relation to the means. In other simulations, the means and variances of the lognormal distributions were specified, using the algebraic relations between the mean and variance of a lognormal distribution and the mean and variance of its normal transform; in some the treatment variances were set equal and in others the variances differed. This corresponds to situations where treatment variances are affected by processes other than those producing the lognormal distribution shape.

Table values of statistics were obtained as follows: F values from Winer (1971), the Dunn-Sidák t statistic from Rohlf and Sokal (1981), and the Studentized Range (Q) values for the Tukey, SNK, Ryan's Q, and Duncan's multiple range tests from Harter (1970). Harter states that his tables for Duncan's test correct errors in the tables provided by Duncan (1955). Steel and Torrie (1960), the most frequently quoted reference for Duncan's test in the biological literature surveyed, reproduce Duncan's (1955) tables. Table values for the Kruskal-Wallis ANOVA were from Hollander and Wolfe (1973). Where possible, exact values for the Nemenyi Joint Rank test from Damico and Wolfe (1987), and for the Steel-Dwass test from Steel (1960) were used. For other values the large-sample approximations of Miller (1981) were used.

RESULTS

Procedures used in ecology

Of the 483 papers from 1982–1984 that compared three or more treatments, 416 (86%) used parametric analyses and 67 (14%) used nonparametric methods. Table 3 summarizes the information on the papers using parametric analyses. Only 7% of these used planned comparisons. In the remaining 385 papers, at least 13 different UMCPs were used to test what appeared to be the same null hypothesis: no differences between pairs of means. In 16 papers, tests designed for other purposes (e.g., all pairwise comparisons) were used to test a control against a number of treatments. With the exception of multiple t testing, most of the tests followed a significant ANOVA or ANCOVA F test. The most commonly used UMCP was the SNK test, followed by Duncan's multiple range test, multiple t testing. Tukey's test, the LSD test, Scheffé's test, and the Bonferroni method. A number of lesser known or newer UMCPs made up 2%, and in the remaining 6% of papers it was not possible to determine which test had been used. The sample of papers from 1986 showed a similar pattern (Table 3), except that the SNK test was less common, and Tukey's test was more frequently used. This may reflect the increased use of the 1981 edition of Sokal and Rohlf, which recommends Tukey's test and does not describe the SNK test.

The average number of treatments compared was ≈5 for the commonly used UMCPs, although in some papers the number of means exceeded 15. It was often not possible to determine the sample sizes used, but where specified they were commonly ≤5. When equal sample sizes were used, the percentage of papers with $n \leq 5$ were: SNK, 32%; Duncan's, 45%; Tukey's, 50%; and LSD, 43%. As shown later, the SNK test produces inflated experimentwise error rates (EERs) when there are groups of treatments with equal means. In 40% of

TABLE 4. Nonparametric unplanned multiple comparison tests used in papers in ecology (1982–1984). Overall test refers to Kruskal-Wallis or Friedman tests.

Test used	Total no. of papers	% of total	Type of ranking	No. after overall test
Multiple Mann-Whitney	24	38	Pairwise	8
Nemenyi Joint-Rank (Dunn's approx.)	12	19	Joint	11
Nemenyi Joint-Rank based on Kruskal-Wallis statistic	9	14	Joint	9
SNK type	8	13	Joint	8
Modified Steel-Dwass	3	5	Pairwise	2
Conover's T test	2	3	Joint	2
Bonferroni-adjusted Mann-Whitney tests	2	3	Pairwise	1
Others	2	3	Joint	2
Unknown	2	3	···	1

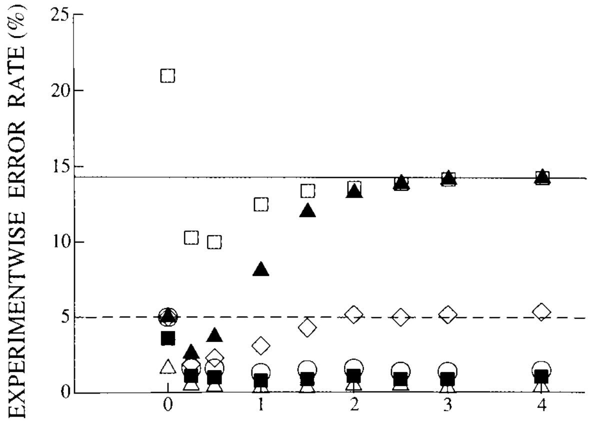

FIG. 1. Experimentwise type I error rates of parametric unplanned multiple comparison procedures (UMCPs) in simulations ($n = 5$). The 6 treatment means were all equal (zero on the abcissa) or in 3 pairs of equal values. The gaps between pairs are shown in units of the population standard deviation (PSD), which was the same for all treatments. ---: nominal 5% significance level, ——: expected error rate for independent t tests. □: Duncan's test. ▲: Student-Newman-Keuls (SNK) test. ◇: Ryan's Q test. ○: Tukey's test. ■: Dunn-Sidák method. △: Scheffé's test.

the papers that used the SNK test, the treatment means apparently fell into two or more groups, e.g., A=B=C ≠ D=E=F. The most commonly cited reference for most tests was Sokal and Rohlf (1969). Zar (1974) and Underwood (1981) were others frequently cited for SNK tests, with the SAS statistical package (Helwig and Council 1979) cited for Duncan's multiple range test, and Steel and Torrie (1960) for Duncan's and the LSD test.

Table 4 shows that the most common nonparametric UMCP was the use of two-sample tests, not corrected for multiple testing. Of the rest, most papers used a joint rank test based on Dunn's approximation or using the Kruskal-Wallis or Friedman statistic, or a stepwise test analogous to the SNK test using the Q distribution (Zar 1974). With the exception of multiple two-sample testing, most authors only applied these tests after a significant overall test (i.e., Kruskal-Wallis or Friedman tests).

Simulations—equal variances, normal distribution.—Parametric tests.—The performance of the commonly used parametric UMCPs is best shown using an example of six treatments each with five observations, with all the tests made at a nominal significance level of 5% (Fig. 1). When there are no real differences between the means (zero on the X-axis), type I errors can occur in all 15 comparisons of means. If all the comparisons were independent, multiple t tests would have an EER of 53.7%. Of the UMCPs studied, only Duncan's test exceeds the 5% EER level. Scheffé's test and the Dunn-Sidák method have EERs below 5%.

When there are three pairs of equal treatments, there are three comparisons of equal means in which type I errors can occur, each of which is independent. The EER for t tests is therefore 14.3%, shown by the solid line in Fig. 1. As the spacing (gaps) between the pairs of means increases from 0.25 to 2 population standard deviations (PSDs), the EERs for Duncan's and the SNK tests approach this theoretical maximum. Tukey's test, the Dunn-Sidák method, and Scheffé's test use the same critical value to test all comparisons, and each remains at a constant experimentwise error level below the nominal rate of 5%. The EER for Ryan's Q test reaches ≈5%. In experiments with two groups of three equal treatments, the same pattern emerges (Fig. 2). The simultaneous tests remain at constant levels <5%, although not as low as in the case of three groups. Ryan's Q test approaches the 5% experimentwise error level, and Duncan's and the SNK tests exceed the 5% level. In this case, the EER for Duncan's test is much greater than for the SNK test, although both remain below the theoretical rate of 26.5% for independent t tests.

The low EERs of the simultaneous tests, especially Scheffé's test, indicate that they would have little power to detect small differences between means within each group of two or three means. However we did not determine their power to do so directly.

The standard deviations of all these results are not presented, but may be calculated from the binomial formula, e.g., 0.10% for an error rate result of 1%, 0.23% for a 5% result, and 0.42% for a result of 20%. A number of runs with three subsets of 3000 experiments were carried out to determine the standard deviation directly, and the results agreed closely with binomial predictions.

These results illustrate the general pattern from our simulations. As Table 5 shows, the EER was always well above the nominal rate for Duncan's test. When-

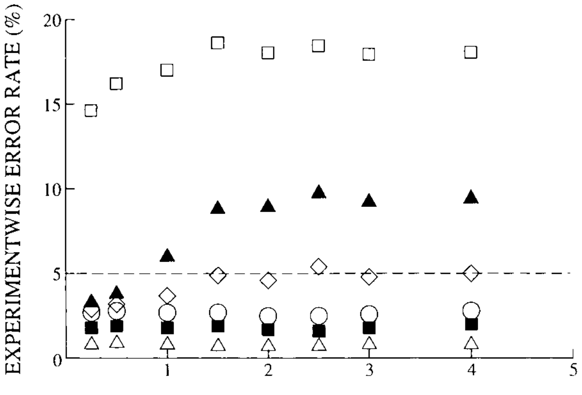

FIG. 2. Experimentwise type I error rates of parametric UMCPs in simulations where treatment means are in 2 groups of 3 equal means. Treatment populations had equal standard deviations (SDS; $n = 5$). Symbols as in Fig. 1.

TABLE 5. Experimentwise type I error rates (EER) of four parametric unplanned multiple comparison procedures (UMCPs) when treatment means fall into groups. Gaps between groups are 3 population standard deviations. Sample size = 5.

Arrangement of means*	No. of groups	Experimentwise error rate (%)			
		SNK	Duncan's	Ryan's	Tukey's
3 equal	1	5.1	9.7	5.1	5.1
1 + 2	1	5.0	5.0	3.4	1.8
5 equal	1	5.0	18.4	5.0	5.0
1 + 4	1	4.6	14.1	3.7	2.9
2 + 3	2	9.8	14.1	4.9	2.5
1 + 2 + 2	2	10.4	10.4	4.1	1.3
6 equal	1	5.0	22.0	5.0	5.0
1 + 5	1	4.9	18.0	4.2	3.7
1 + 1 + 4	1	5.2	14.4	3.6	2.5
2 + 4	2	9.7	18.0	5.0	3.2
3 + 3	2	9.2	17.9	4.8	2.6
1 + 2 + 3	2	9.5	13.9	4.5	1.8
2 + 2 + 2	3	13.7	13.7	4.9	1.3

* Number, and size, of groups of equal means. For explanation see Table 1 footnote.

ever there was more than one group of equal means and the gaps between groups were large, the SNK test also exceeded the nominal rate; the severity of the problem depends on the number of groups rather than the total number of means compared. For Duncan's test, the EER depends on both the number and size of the groups. The EERs of tests are not greatly affected by sample size. In fact, in the range n = 3–9, which occurs commonly in the biological literature, the changes in error rate were too small to detect reliably.

Most ecologists used UMCPs only when the ANOVA F test was significant (Table 3), probably because most texts state that UMCPs would only detect differences already indicated as present by the ANOVA, rather than to "protect" EERs. Protection is not necessary for those tests that control EERs, but might be expected to reduce the error rates for the SNK and Duncan's tests. The protection works when there are no real differences between any means, as the EER cannot exceed the 5% error rate of the F test. Our simulations showed, however, that when there are differences between groups of means, protection by the F test has very little effect. As the spacing between groups increases, the power of the F test increases, so that F is significant in most experiments, and the EERs of the protected SNK and Duncan's tests become equal to those of the unprotected tests. In our results with six means in three pairs and a sample size of five, this occurred when the pairs were one standard deviation (SD) apart.

TABLE 6. Effect of unequal sample sizes on error rates (%) of ANOVA and Tukey's test when variances are heterogeneous (3 treatments, variances = 1, 1, 10).

Sample sizes	Box's coefficient	Bias ratio	Error rate of ANOVA F	Error rate of Tukey's
9, 9, 9	1.1	1.0	7.8	7.4
9, 9, 7	1.2	1.2	11.6	11.2
9, 9, 5	1.3	1.4	17.2	17.5
9, 9, 3	1.4	2.0	26.8	27.4
3, 9, 9	0.8	0.8	3.7	3.4
6, 6, 9	0.9	0.8	4.1	3.3

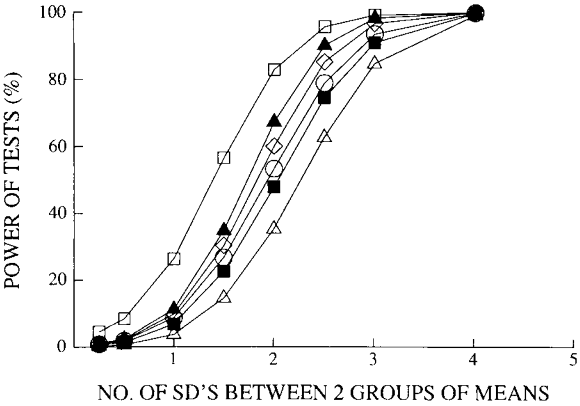

FIG. 3. The power of parametric UMCPs to separate means in 2 groups of 3. Symbols and conditions of the experiments as in Fig. 2.

The power of tests (percent of correct decisions when real differences occur) can be illustrated in the case of two groups of three means, because the real differences between means in different groups are the same for each comparison (Fig. 3). The powers of those tests with EERs ≤5% are less than for the SNK and Duncan's test, simply as a consequence of higher type I error rates for the SNK and Duncan's test. Higher power could be achieved using the other tests by setting an EER of, e.g., 10%. The differences between tests are largest for real differences of about two standard deviations (Fig. 3). Here the power of Ryan's Q test is ≈60%, and the power of Tukey's test is ≈53%, but the power of Scheffé's test is only 35%. The relative power of the tests will vary with the arrangement of the means into groups and the differences between groups, but the ranking of the tests will remain the same. The increased power of Ryan's Q test relative to the other tests which control the EER will be greatest when there are many groups and many means.

Sample size has an important effect on the power of Tukey's and Ryan's Q tests (Fig. 4A, B). While these tests differ in power by only ≈7% at most when n = 5 (Fig. 3), the power of each may be increased by up to 40% by raising the sample size to nine. Sample size affects the power of other tests in a very similar way.

Nonparametric tests.—When means were in groups the EERs of the nonparametric tests altered with the spacing of the groups in a similar way to Tukey's and Ryan's Q tests (Fig. 5A). Results are shown for a sample size of 9 because the Steel-Dwass test cannot be used for six treatments with a sample size <6.

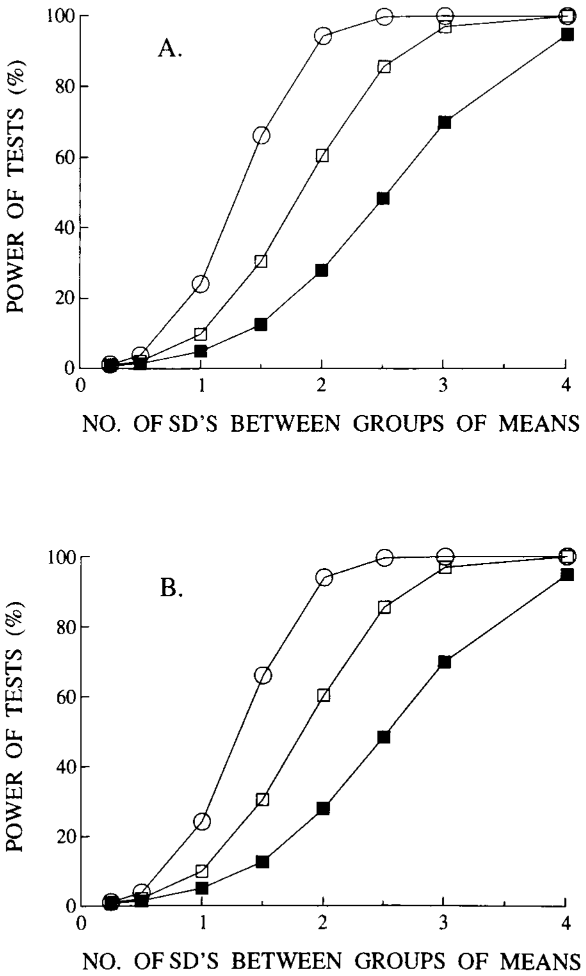

FIG. 4. Effects of sample size on the power of unplanned multiple comparison procedures (UMCPs). Six treatment populations with equal standard deviations. Means in 2 groups of 3. (A) Tukey's test. (B) Ryan's Q test. Symbol indicates sample size. ■: $n = 3$. □: $n = 5$. ○: $n = 9$.

All the nonparametric tests are more conservative than their parametric equivalents for normally distributed data, which indicates that they would have less power to detect differences between means within each group. The Nemenyi Joint Rank test becomes extremely conservative as the gap between groups increases because the joint ranking of a group of treatments with other very different treatments reduces the relative differences between rank sums within a group. This is an important disadvantage of this test, and one of the reasons why the stepwise version, which requires re-ranking within each group, is much more useful. The stepwise Steel-Dwass Ryan test is extremely conservative when the spacing between groups is small, because two separate criteria are applied in comparing treatments. Rank sums are used to determine significance, but the treatment medians are used to determine which treatments are farthest apart and therefore compared first. It is possible that the medians of two of the treatments may be far apart but their rank sums not very different. If so, stepwise testing would stop at this pair, and other treatment pairs with very different rank sums would be declared nonsignificant.

The relative power of the nonparametric tests are shown in Fig. 5B, for six treatments in two groups and a sample size of nine. Note that these power comparisons are for normally distributed data. The nonparametric tests were always less powerful than Ryan's Q or Tukey's tests. The Steel-Dwass Ryan test has very low power because it is so conservative, so that the most powerful of these tests is the Joint-Rank Ryan test. The two simultaneous tests have similar power in this situation, but simulations with three treatments

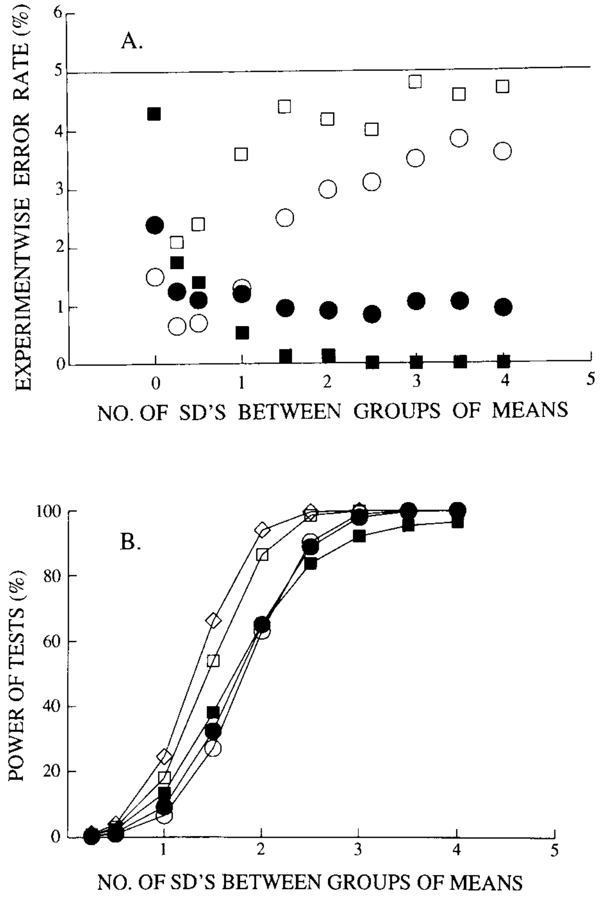

FIG. 5. Error rates and power of nonparametric UMCPs, when data are normally distributed. Six treatment populations with equal standard deviations, and means in 2 groups of 3, as in Figs. 2 and 3, but sample size = 9. (A) Experimentwise error rates as spacing between groups increases. Horizontal line indicates nominal 5% significance level. (B) Power of tests to separate groups, as compared to the parametric Ryan's Q test. ■: Nemenyi Joint-Rank test. □: Joint-Rank Ryan test. ●: Steel-Dwass test. ○: Steel-Dwass Ryan test. ◇: parametric Ryan's Q test.

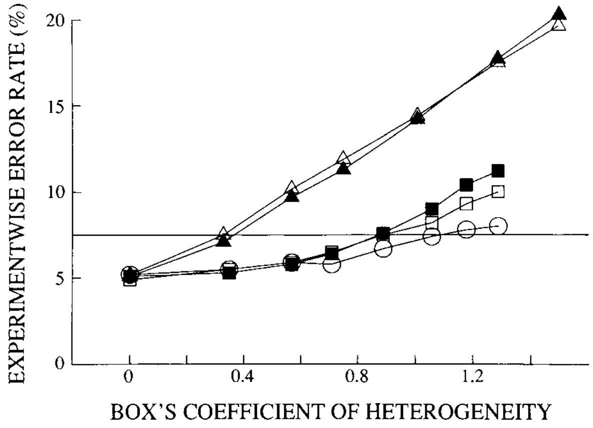

FIG. 6. The effect of variance heterogeneity on Tukey's test, with three treatments, means equal. The nominal significance level is 5%, and Bradley's (1978) liberal criterion of robustness is indicated by the line at 7.5%. ■: n = 3, 3, 3. □: n = 5, 5, 5. ○: n = 9, 9, 9. ▲: n = 5, 5, 3. △: n = 9, 9, 5.

showed that the Steel-Dwass test was always more powerful. As the Steel-Dwass test uses the data in each pair of treatments separately, it should perform best where the sample size is large compared to the number of treatments.

Simulations—unequal variances, normal distribution.—When the treatments do not differ, the EERs of the stepwise versions of tests are the same as the simultaneous tests, so the effects of unequal variances when treatments are equal are described below for Tukey's test, the Nemenyi Joint Rank test, and the Steel-Dwass test only. Tukey's test is affected in the same way as the ANOVA F test with respect to Box's coefficient of heterogeneity and his bias ratio (Rogan et al. 1977; our simulations). High values of Box's coefficient can occur because one variance is fairly large while the rest are small, or as a result of a very wide range of variances. For example, when the variances are 1, 1, and 10 or 1, 1, 1, 1, 1.5, and 7, Box's coefficient has the same value as for variances of 1, 16, and 80. As a narrower range of variances would seem more likely to occur in practice, this is the situation we illustrate, and for which ecologists should check their data.

With three treatments and equal sample sizes, the EER of Tukey's test increased slowly but markedly as variance heterogeneity increased; it rose to ≈10% when one variance was very large compared to the others (Fig. 6). Bradley (1978) suggested, as a liberal definition of robustness, that the type I error rate should not exceed 1.5 times the nominal rate. This level is shown by the horizontal dashed line in Fig. 6. Further simulations with six treatments have shown that high EERs can occur in many ways when there are many treatments, namely: if one (or a few) of the many variances is large, if larger variances occur in one subgroup of treatments, or if there is a wide range of treatment variances. The effects are also more severe with many treatments. Note that the error rates reported here are for the set of all pairwise comparisons in an experiment. For some particular comparisons, the error rate may be much greater.

When sample sizes were not equal, and small samples were associated with larger variances, the error rate rose much more rapidly as the variance heterogeneity increased (Fig. 6). For example, when population variances were 1, 1, and 10, Box's coefficient was 1.1 and the error rate was 8.3% with all sample sizes set at 5, whereas with sample sizes of 5, 5, and 3, Box's coefficient became 1.3 and the error rate was 18.1%. Table 6 shows how the EER of both the ANOVA and Tukey's test vary with the bias ratio and sample sizes. Bias ratios >1.1 (when smaller samples have larger variances) produce highly inflated error rates.

The nonparametric tests were investigated only for equal sample sizes. The effects of variance heterogeneity, where one variance is larger than the others, are shown in Fig. 7A, B. The EERs were much less inflated

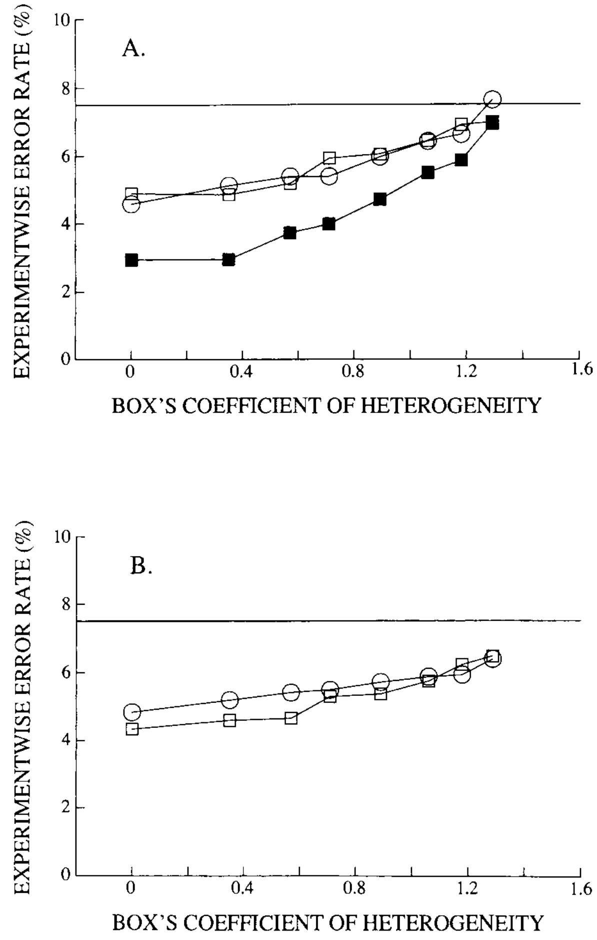

FIG. 7. The effect of variance heterogeneity on nonparametric UMCPs. (A) Nemenyi Joint-Rank test. (B) Steel-Dwass test. Symbol indicates sample sizes as in Fig. 6.

than the EER of Tukey's test. The Steel-Dwass test was slightly more robust than the Nemenyi Joint Rank test. However both tests are less robust when two or more variances are large, and unequal variances are expected to have more effect when sample sizes differ.

Fig. 8A shows the frequency distribution of estimates of Box's coefficient calculated from data in the ecological literature. Many of the papers analyzed several attributes measured under the same treatments, so that not all of the values are independent, and these results must be treated with caution. The distribution is highly skewed, however, with ≈16% of values >1, and a long tail of extreme values. A coefficient of 1 would lead to EERs >7.5% even if sample sizes were equal and there were only three treatments. The non-independence would probably have reduced the spread of our data, so that the real situation may be worse. As sample sizes become more unequal, the bias ratio becomes the major factor inflating EERs, and bias ratios >1.1 produce highly inflated EERs. The distribution of estimated bias ratios in the literature survey is shown in Fig. 8B. Although the non-independence problem is more severe here, and may reduce the spread of the data, <6% of the values are >1.1, so that high bias ratios appear to be rare in practice.

Simulations—lognormal data.—Biological data often follow a lognormal distribution as a result of underlying multiplicative effects. These cause the distribution of measurements under each treatment to have a lognormal shape, and also the variances under each treatment to be related to the means. A log transformation renders the distribution normal (i.e., on a log scale) and the variances equal. The simulations in Table 7 have been arranged to show these two effects. In each simulation there were six treatments in two groups, the second group having larger means (by either 0.6 or ≈2 PSDs) than the first.

In simulation A the distributions are lognormal, but the variances have been set to be equal to investigate the effect of the lognormal shape alone. This appears to reduce the EER of the parametric tests, so that they are conservative. In simulation B, the means and variances of the second group are larger (as expected in lognormal biological data). A log transformation would result in treatment populations with equal variances and means 2 PSDs apart, but the simulated data were analyzed without transforming. The EERs of the parametric tests are inflated, and their power is dramatically reduced, but the nonparametric tests are not affected. Simulation C is similar to B, but the means and variances of each treatment have been chosen so that, as in simulation A, the average variance of the two groups is 4 and the difference between the groups is ≈2 PSDs. The EERs of the parametric tests remain somewhat inflated, but less than in B. In simulation D the data are as in B but have been transformed to the log scale before analysis. The parametric tests have EERs much closer to 5% and much better power than the nonparametric tests.

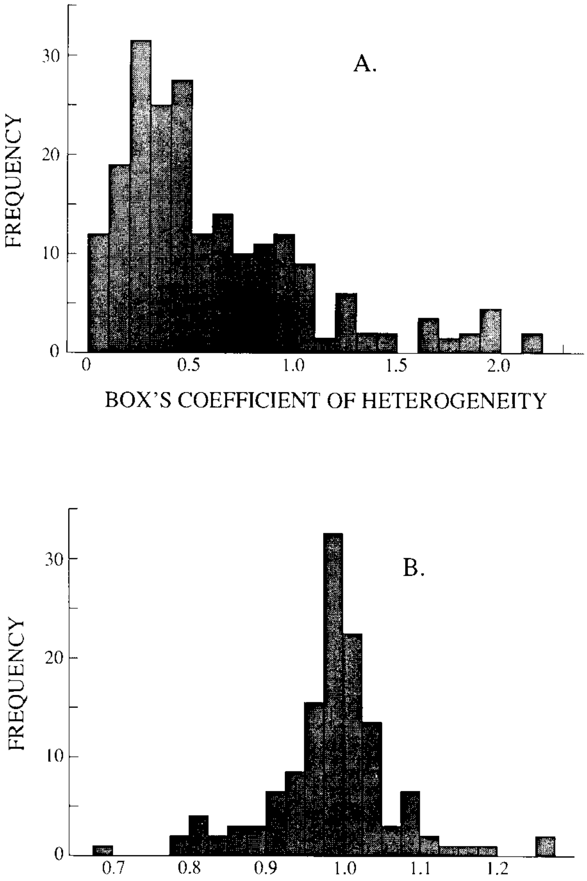

FIG. 8. Frequency distributions of estimated values from experiments in ecology. (A) Box's coefficient of variance heterogeneity. (B) Box's bias ratio.

These results illustrate that, for lognormal data, nonparametric tests perform well if variances are equal either on the original scale of measurement or on some transformed scale. They also show that the most important problem of lognormal data in biology is the unequal variances in such data, and demonstrate the value of graphing means against variances as a check of whether a log transformation should be used. On transformed data the parametric tests behave well. Finally, these data confirm that the Joint-Rank Ryan test is the most powerful of the nonparametric UMCPs.

Discussion and Recommendations

Assumptions

The parametric statistics used for the ANOVA and for comparisons are often claimed to be robust to the assumptions on which they are based (e.g., Hays 1981). Robustness, however, is a relative term (Bradley 1978), and some assumptions, such as that sampling is random and that observations are independent, are essential. The consequences of various kinds of non-

TABLE 7. Effect of lognormal data on parameteric and nonparametric unplanned multiple comparison procedures (UMCPs). In each simulation there were 6 treatments with means in 2 groups, $n = 9$. EER = experimentwise Type I error rate. PSDs = population standard deviations.

Simu-	Variances		Spacing		UMCPs*					
lation	Group 1	Group 2	(PSDs)	Statistic	Tukey	Ryan	S-D	N-J-R	S-D-R	J-R-R
A	4	4	2	EER (%)	1.9	4.1	1.1	0.3	3.0	3.9
				Power (%)	89	94	65	65	62	87
B	5.7	38.1	0.6	EER (%)	7.5	12.5	1.2	0	3.7	4.7
				Power (%)	28	32	64	65	62	86
C	1.7	6.3	1.8	EER (%)	5.7	9.3	1.1	0	3.2	4.5
				Power (%)	81	88	65	65	63	86
D	5.7	38.1	0.6	EER (%)	2.4	5.0	0.9	0.1	3.0	4.2
				Power (%)	90	94	65	65	63	86

* S-D: Steel-Dwass test, N-J-R: Nemenyi Joint-Rank test, S-D-R: Steel-Dwass Ryan test, J-R-R: Joint-Rank Ryan test.

independence are discussed by Kenny and Judd (1986). Ecologists should consider how correlations between observations may arise in their data, and describe what precautions have been taken to minimize possible correlations.

The effect of non-normal distributions has been investigated in many studies (e.g., *F* statistic: Gayen 1950; *t* statistic: Boneau 1960, Ratcliffe 1968, Posten 1984; *Q* statistic: Ramseyer and Tcheng 1973, Brown 1974). It would appear that, if sample sizes and variances are close to equal, only severe non-normality (e.g., bimodal distributions or extreme skewness) or the presence of outliers is a problem for parametric one-way ANOVAs and UMCPs. We have shown that data that follow a lognormal distribution as a result of multiplicative processes can result in inflated EERs for parametric UMCPs, but the major problem is the resulting unequal variances (discussed later in this section). Milligan et al. (1987) have shown that non-normality also affects the computations for unequal cell sizes in factorial ANOVA designs. Thus ecologists should consider what processes give rise to the variation in their data, especially when sample sizes differ, and whether these might produce discontinuous effects (resulting in outliers) or a multimodal distribution. An ecological example of processes producing a bimodal distribution might be the growth rates of animals where an undetected parasite affects growth markedly. The distribution of growth rates in the population would then have a mode for parasite-free animals and another for infected animals. This is similar to a case described by Bradley (1977) that affected tests severely.

Tests for normality are rarely considered in biology, probably because of the small sample sizes used. Ecologists should at least take simple, commonsense precautions to check for non-normality, particularly as the processes which produce non-normal distributions may also result in variance heterogeneity. Such precautions are discussed in texts on exploratory data analysis (Tukey 1977, Hartwig and Dearing 1979, Hoaglin et al. 1983; see also Miller 1986, Dunn and Clark 1987). Some non-normal distributions can be detected from the relationship between sample means and variances or residuals. Alternatively, one can use a probit plot of the data in each treatment, where sample sizes are large enough (say, >8), or a probit plot of the residuals in all treatments if the variances and shape of the distributions can be assumed to be the same (Winer 1971, Miller 1986).

A transformation to a different scale of measurement will often produce normality of distributions within treatments (e.g., Snedecor and Cochran 1980, Sokal and Rohlf 1981; see Underwood [1981] for application of transformations to ecology). We have demonstrated that where multiplicative processes produce lognormal data, transforming to a log scale is the best strategy, as the more powerful parametric UMCPs control the EER. The interpretation of analyses of transformed data, however, needs considerable care (see below in this section). Nonparametric tests should be considered in situations where it is impossible to tell if the data are close to normal.

Outliers may affect even tests robust to the usual normality assumptions (Dunnett 1982), and therefore some objective "check" for outliers is useful. Dunn and Clark (1987) discuss significance tests as checks for outliers (assuming normal distributions), and Chambers et al. (1983) and Hoaglin et al. (1983) describe "outlier cutoffs" based on certain ways of summarizing batches of data. These summaries can be very usefully presented as box plots (or schematic plots), which provide visual checks for outliers. McGill et al. (1978) provide some useful refinements to the box plots technique. Identifying outliers allows the ecologist to check for errors. Underwood (1981) correctly cautions against the routine removal of "unsatisfactory" data unless they are clearly impossible from the nature of the experiment. Techniques for removing outliers are usually based on winsorizing (see Winer 1971) and are discussed in some detail in Hoaglin et al. (1983).

Relatively few studies have investigated the assumption that all treatments have the same type of distribution. Error rates for the t test can be inflated when samples are small or unequal and the treatment distributions are skewed to different extents or directions (Boneau 1960). This might occur if the biological processes producing variation differed between treatments. One-tailed tests were especially affected. However, Ramseyer and Tcheng (1973) found that the error rate of Tukey's test, taken over all the pairwise comparisons, was little affected by distributions with different shapes, even when they were skewed in different directions. Neither transformations nor the standard nonparametric methods can bypass this problem, although robust nonparametric methods have recently been proposed that rely on the distributions being symmetrical rather than the same (Fligner and Policello 1981, Rust and Fligner 1984).

Early studies (e.g., Box 1954, Boneau 1960, Ramseyer and Tcheng 1973) stressed the robustness of the F, t, and Q statistics to heterogeneity of variance. Later studies (Games and Howell 1976, Rogan and Keselman 1977, Rogan et al. 1977, Dunnett 1980*b*, Wilcox et al. 1986, our data) have emphasized the severe effects of more unequal variances on the ANOVA and Tukey's test. Yet the commonly used textbooks contain only vague statements such as "the consequences of moderate heterogeneity of variance are not too serious" (Sokal and Rohlf 1981:408). An important point is that in factorial experiments where variances are heterogeneous, treatments cannot be pooled across one factor to make comparisons between levels of another factor. The standard computational routines for unequal cell sizes in factorial ANOVA designs are also not robust to variance heterogeneity (Milligan et al. 1987).

As with non-normality, the problems of variance heterogeneity for one-way designs are more severe with unequal sample sizes. When all treatment means are equal, our simulations show the EERs of Tukey's test are very similar to those for the ANOVA F, and closely related to Box's coefficient of heterogeneity and bias ratio. Our sample of the ecological literature suggests that the bias ratio is seldom extreme, but the coefficient is often high enough to seriously increase error rates. For equal sample sizes and normal distributions the nonparametric UMCPs appear more robust to variance heterogeneity than their parametric counterparts, although seriously inflated EERs still occur. The effects of combinations of variance heterogeneity and non-normal distributions have not been studied in detail.

Some patterns of variance heterogeneity can be detected with the methods for exploratory data analysis referred to above (e.g., relationships of means and variances). More commonly, significance tests of the equality of variances are used. Ecologists should be aware, however, that if preliminary tests of variances are used to determine how to analyze the data, this will change the error rate of the overall analysis. Tests to detect heterogeneity of variance are useful only in some cases (R. W. Day and G. P. Quinn, *personal observations*). The most powerful variance check that is robust to non-normality, the Levene test (Snedecor and Cochran 1980) using medians instead of means (Brown and Forsythe 1974*a*), provides very little protection for Tukey's or Ryan's Q tests unless sample sizes are equal and >9. Wilcox et al. (1986) studied the effect of protecting the ANOVA F test using the Levene-median test, and reached the same conclusion. We are not aware of any other studies of this type.

Unless distributions are known to be normal, Cochran's, Bartlett's, or Hartley's tests (Winer 1971, Underwood 1981), or any F test for variances, should not be used, as they all have inflated type I error rates for non-normal distributions (Box 1953, Snedecor and Cochran 1980, Rivest 1986). Our results for simulated normal data have shown that, for a given range of variances in the data, the most severe effect on the EER of an UMCP occurs when there is one deviant large variance. Cochran's test is the most powerful for detecting this situation (R. W. Day and G. P. Quinn, *personal observations*). When several variances are large, a much greater range of variances is required to produce the same effect on UMCPs. Other variance checks are then more powerful, but such situations seem less likely to occur in practice. We therefore recommend Cochran's test if data are normal. However, Cochran's test only affords protection when sample sizes are fairly balanced and >5 if three treatments are compared; slightly larger sample sizes will be needed with more treatments (R. W. Day and G. P. Quinn, *personal observations*). With unequal sample sizes, the effect of unequal variances is more severe. An F test of the variance of the smallest sample vs. the average variance of the others may be useful (R. W. Day and G. P. Quinn, *personal observations*) if used at a 25% significance level to ensure detection of cases where the smallest sample has a large variance, as this has the most severe effect on UMCPs. This F test will not detect other kinds of variance heterogeneity, so that it should only be used alongside Cochran's test.

Many ecologists appear to transform the data when variances are found to be significantly different. We have shown how effective this is when using UMCPs on suitable lognormal data. Snedecor and Cochran (1980) pointed out that this will change the distribution of the data, and the additivity of effects in factorial designs, as well as the variances. In our example, the log-transform changed the distribution from lognormal to normal and multiplicative effects to additive effects, as well as made variances equal. Snedecor and Cochran (1980) described a technique to assess simultaneously the different effects of transformations in factorial ANOVAs. Although changes to the distribution would appear to have relatively minor effects (see above in this section), the additivity problem affects the biological interpretation of interactions. Other problems arise if means have to be converted back to the original scale to draw meaningful biological conclusions. In this regard, Games and Lucas (1966:326) conclude that "the use of a clearly interpretable scale of measurement certainly should be the dominant consideration." While transformations are an important method of overcoming variance heterogeneity (see Underwood 1981), the problems outlined above should be kept in mind. The alternative is to use UMCPs and versions of ANOVA which are robust to variance heterogeneity; although standard nonparametric tests are better than parametric ones in this regard, they are not fully robust to variance heterogeneity (see next section, below).

Alternatives to the ANOVA

Parametric.—Our simulations and those of Kohr and Games (1974), Tomarken and Serlin (1986), and Wilcox et al. (1986), indicate that the standard ANOVA F may be severely affected by variance heterogeneity. Unweighted-means ANOVA is even more affected (Milligan et al. 1987). As variance checks are not useful at small sample sizes, robust versions of ANOVA may be warranted. At least four alternatives to one-way ANOVA, the W test (Welch 1951), the BF test (Brown and Forsythe 1974*b*), the Fisher-Pitman permutation test (Still and White 1981), and the two-stage Bishop-Dudewicz procedure (Bishop and Dudewicz 1978, 1981) have been proposed. The permutation test, however, has been shown to require equal variances when used as a test of equality of means (Boik 1987) and is unsuitable as a robust alternative. Wilcox et al. (1986) found that for a nominal significance level of 5% and equal sample sizes >6, W appears to control the type I error at <7.5% when variances are unequal, whereas BF sometimes does not. For extremely unequal sample sizes (max/min ratio = 4), W may be too liberal; in fact neither test is robust to some patterns of variance heterogeneity, especially for many treatments ($m > 4$). However both Kohr and Games (1974) and Tomarken and Serlin (1986) found that W is robust for three or four treatments when the max/min ratio of sample sizes was 3. Clearly, very unequal sample sizes should be avoided. While the results of Brown and Forsythe (1974*b*) and Dijkstra and Werter (1981) suggest BF may sometimes be more powerful for $n < 6$, we recommend W on the basis of better control of type I errors.

The Bishop-Dudewicz procedure is a two-stage test in which a random sample of observations is taken from each treatment group and, after some calculations, based partly on required power, the number of additional observations needed can be determined. The null hypothesis of equal means is then tested with a modified F statistic for which approximations or tables are available (Bishop and Dudewicz 1978, Wilcox 1986). Limited simulation data (Bishop and Dudewicz 1981) suggest its robustness and power are as good as the W test. The advantages of the Bishop-Dudewicz procedure are that it provides excellent control of power of the test and is suitable for factorial designs. Its applicability to ecology, however, appears restricted because few field ecologists have the resources for two-stage sampling (hence the rarity of pilot studies) and the sample size needed at the first stage is often large (>15).

The loss in power, for normal distributions, of the robust alternatives to one-way ANOVA and tests for comparisons, relative to the standard methods, is typically <5% in simulations, so that in cases where the variance checks are ineffective the standard methods could simply be abandoned in favour of the robust methods. No robust alternative is available for nested and factorial ANOVA designs because variance-pooling assumptions are involved (see discussion in Milligan et al. [1987]).

The above applies to normal distributions. The robustness of these tests to variance heterogeneity combined with non-normal distributions is not clear (Levy 1978, Clinch and Keselman 1982, Tan and Tabatabai 1986, Wilcox 1986). We suggest that, until more information is available, these robust alternatives should be avoided when distributions are likely to be non-normal.

Non-parametric.—The Kruskal-Wallis test for one-way designs and the Friedman test for two-way designs without replication do not require that the distributions be normal. However, except for a difference in medians, the distributions must be identical in all the treatment populations compared, either for the original data or for some transform of the data (Hollander and Wolfe 1973, Conover 1980). It follows that, while the variances need not be equal in the raw data, there must be a suitable transform to stabilize the variances. If transforms are appropriate (see above, Assumptions), parametric analyses of transformed data would be more powerful. The standard nonparametric tests should not be used as a simple means to avoid the problem of unequal variances, as some authors of the papers surveyed appeared to do. In our simulations, the EER of the Kruskal-Wallis test increased as variances became more unequal, although not as rapidly as for the ANOVA F test. Rust and Fligner (1984) have described a robust version of the Kruskal-Wallis test which requires only that the distributions be symmetrical, not identical.

Although these nonparametric procedures do not allow the elegant partitioning of the sources of variation provided by the ANOVA, significance tests of main effects and interactions in factorial designs can be carried out (Bradley 1968, Patel and Hoel 1973, Groggel and Skillings 1986; see also Zar 1984).

Choice of error rates for comparisons

A consideration of type I error rates is important in determining which of the many UMCPs should be used, and leads to a discussion of whether biologists should frame their questions to use more powerful techniques, such as planned comparisons. Only two choices of type

I error rate for comparisons seem reasonable: a per-comparison error rate or an experimentwise error rate (EER). Which of these should be used depends to a large extent on the objectives of the experiment.

Many texts point out that if the experiment is done without any plans as to how to examine the results, and the results suggest a comparison of means or combinations of means to the researcher, then an EER should be used (e.g., Snedecor and Cochran 1980). The analysis is picking a result from among the possible ways to compare the means, and the EER is then the probability of finding one or more such results that are significant by chance.

Similarly, if the experiment is conducted in order to make a collection of pairwise comparisons, i.e., to explore and interpret how the treatments differ, an EER also seems appropriate, as recommended in a number of texts (e.g., Snedecor and Cochran 1980, Keppel 1982). This is both because the comparisons are not orthogonal, and because one is "picking winners" from the possible comparisons, such that conclusions will be based on the differences which are significant. However, opinions on this topic vary. Some authors (e.g., Carmer and Swanson 1973, Carmer and Walker 1982) argue that each pairwise comparison is of individual interest, and should be tested using a per-comparison error rate. One form of this argument is that if three different researchers each compared two treatments, the per-comparison error rate would be used, and therefore one researcher who compares three treatments (three comparisons) should not be forced to use a less powerful test. We do not subscribe to this argument. The single researcher does not have three independent sets of data for the three comparisons. He/she does, however, have a better estimate of the variance within groups than if only two treatments were studied, provided the treatment variances are homogeneous. This estimate of variance with more degrees of freedom allows for a more powerful test, given the EER used.

Another common argument against using an EER for unplanned comparisons is the lack of power of the tests. Sokal and Rohlf (1981:243) state that this loss of power "is the price one pays for testing unplanned comparisons." Power can obviously be increased by using tests with less stringent type I error rates than experimentwise. We have already argued against using a per-comparison error rate and, in agreement with Ryan (1959), Scheffé (1959), and Miller (1981), we find tests with an indeterminate error rate, i.e., one that can exceed the nominal level by an unknown amount, unacceptable. One of the bases of using statistics to test hypotheses is that a reliable probability statement concerning type I errors can be associated with any decision. Those tests with an indeterminate error rate (e.g., Duncan's and the SNK) do not allow this.

The power of UMCPs, as with other statistical tests, can be increased by increasing the sample size. Indeed, our results show that increasing the sample size slightly will increase power far more than choosing, say, the SNK rather than Tukey's test. Morley (1982) pointed out that the power of all the UMCPs decreases as the number of treatments increases, so that small focused experiments are better than larger "hope-something-shows-up" experiments. If sample sizes are unavoidably small, we suggest two alternative strategies. First, one can decide to risk more type I errors, at a known rate, by increasing the nominal significance level used with an experimentwise test—there is nothing "sacred" about 5%. Snedecor and Cochran (1980) suggested that those worried about power should use a 10% or 25% EER. If a more sophisticated assessment of the seriousness of type I and type II errors can be made, the Waller-Duncan k-ratio t tests might be appropriate (but see Unplanned Comparisons, below). Second, one might consider a category of results (or decision rules) advocated by Keppel (1982) called "suspend judgement"—make no decision on rejecting the null hypothesis when the test statistic falls between the critical values at the per-comparison and experimentwise levels. In a strict sense, however, this conclusion should apply to any non-significant result. Although often not feasible, the ideal approach is for biologists to design their experiments and analyses with the kind, and size, of effect they wish to detect in mind. This idea is far from original, and includes both planning comparisons and a priori power analysis (Winer 1971, Cohen 1977, Underwood 1981), although the application of power analysis to multiple comparison procedures is rare.

Planned comparisons

Planned comparisons of means or combinations of means have definite advantages (Winer 1971, Snedecor and Cochran 1980, Sokal and Rohlf 1981, Keppel 1982). They test a specific hypothesis about the treatments which flows from the logic of the experiment and can be stated in advance. Often several planned comparisons are made. If they are orthogonal, and the researcher is not "picking a winner," then each comparison provides a separate answer to a separate biological question, and therefore a per-comparison error rate is appropriate. This means that planned comparisons are more powerful tests than unplanned comparisons. Planned comparisons also facilitate the use of a priori power analysis. The reason few ecological papers used planned comparisons may be simply because most experiments in ecology are exploratory, but we suspect planned comparisons, in particular trend analysis (described later in this section), could have been useful in many papers. Ecologists should consider whether they can ask meaningful, focused questions, and use planned comparisons in their analyses to take advantage of these methods. Rosenthal and Rosnow (1985) discuss parametric planned comparisons extensively, with numerous examples.

Unfortunately, the relevant and interesting comparisons in an analysis are often not all orthogonal. In this case, the hypotheses tested by the comparisons may be interrelated in some way; for example, if one tests whether A differs from B and whether A differs from C, then the answers both depend on whether the sample for A is unusual or not. Winer (1971) argued that the meaningfulness of the comparisons in the context of the experiment is more important than their orthogonality. Where a number of comparisons are not orthogonal to each other, however, the significance levels should be adjusted for these comparisons to control the EER (Sokal and Rohlf 1981). The Dunn-Sidák method provides the most powerful experimentwise test for a predetermined number of such comparisons (Miller 1981, Sokal and Rohlf 1981). To maximize the power of the tests the number of non-orthogonal tests should be small. The researcher should also realize that the information from each non-orthogonal comparison is not independent, and interpret them cautiously to avoid ambiguities. When the number of planned comparisons exceeds the number of degrees of freedom, some must be non-orthogonal. Keppel (1982) suggests these can be managed by making some tests only if others are significant.

A useful branch of planned comparisons is trend (or response-curve) analysis, which is used when the treatments are levels of some quantitative factor (e.g., density, depth, temperature). A series of orthogonal planned comparisons can be designed to answer such questions as whether the means show a trend with increasing levels of the factor, or whether a threshold exists beyond which the factor has a greater (or lesser) effect. These questions may be more informative than UMCPs; in fact, Petersen (1977), Little (1978, 1981), Baker (1980) and others have argued that UMCPs should never be used for agricultural experiments where trend analysis is appropriate. Dawkins (1983) suggested that biologists do not use trend analysis because it is not in most elementary texts, and it appears more complex than it is. Parametric methods are extensively reviewed by Mead and Pike (1975) and presented in Keppel (1982). They are described under "orthogonal polynomials" in commonly used texts (e.g., Winer 1971, Sokal and Rohlf 1981), but these often deal with complex polynomials, which would seldom be required for comparisons, and do not cover the methods needed when the levels of the factor are unequally spaced (see Robson 1959, Kendall and Stuart 1967, Keppel 1982). Keppel (1982) discusses traps in the use of parametric trend analysis (e.g., problems with interpolation, extrapolation, and transformed data). There are also nonparametric tests for orderings of the treatments. These do not require numerical values for levels of the factor. Hollander and Wolfe (1973) discuss these tests in one-way and two-way designs, and Page's test of trend is described in Sokal and Rohlf (1981).

Variance heterogeneity may seriously affect planned comparisons (Keppel 1982), so that a test of the variances involved in a planned comparison would seem appropriate. Such tests, however, would have very low power with the small sample sizes commonly used, and it may be best to always use a test robust to unequal variances. Studies by Mehta and Srinivasan (1970), Wang (1971), Davenport and Webster (1975), Kohr and Games (1977), and Best and Rayner (1987) all suggest that the Welch (1938) *t* test, generalized with Satterthwaite's degrees of freedom (see Winer 1971), is the most powerful and robust test. Best and Rayner (1987) state that the loss in power is so small for $n > 5$ that the Welch *t* test should always be used in place of Student's *t* test.

Unplanned comparisons

Parametric tests (see Table 8).—1. *Pairwise comparisons—equal sample sizes and variances.*—A number of UMCPs commonly used in ecology for comparing all pairs of means do not control the EER at or below the value required. Our simulation results show that Duncan's test uses an indeterminate type I error rate which is between comparisonwise and experimentwise; i.e., the EER is not controlled at the chosen significance level. This is stated in a number of texts (e.g., Scheffé 1959, Steel and Torrie 1960, Winer 1971), but does not seem to be acknowledged by ecologists who use it. The error rate is often much higher than the significance level chosen. Our simulations also demonstrate that the SNK method controls the EER at the chosen significance level only if the treatments do not fall into groups. This problem has been noted by Ryan (1960), Petrinovich and Hardyck (1969), Snedecor and Cochran (1980) and Klockers and Sax (1986), but is apparently not widely known. Commonly used references for the SNK test either imply that it does control EERs at the nominal level (e.g., Zar 1974; but see also Zar 1984) or discuss the test in the context of EERs without describing how it performs (e.g., Sokal and Rohlf 1969, Underwood 1981). We suspect that many biologists using the SNK test assume that the EER is at or below the nominal significance level. When treatments fall into groups, which may occur if there are >3 treatments, the EER varies, becoming higher if there are many treatment groups and the groups are well spaced.

We therefore argue that these two tests (Duncan's and SNK), with their unknown EERs, are not appropriate. It is impossible to determine how often inflated error rates for these tests occur in ecology, because the true means of treatments are not revealed by the results. Our survey suggests that it may occur often, because means often appear to fall into two or more groups, and large *F* ratios indicate that in many cases large differences occur. Making these tests conditional on a significant ANOVA *F* test (i.e., "protected") does not substantially alter the error rates unless the differences between treatment groups are very small.

Carmer and Swanson (1973) and others have claimed

TABLE 8. Summary of suggested or commonly used parametric tests for pairwise comparisons and contrasts (comparing combinations of means). C = per-comparison (comparisonwise); E = experimentwise; I = indeterminate; NS = not simultaneous.

Test	Usage and assumptions	Type I error	Confidence intervals	Recommended/remarks
Single-df orthogonal contrasts	pairwise/contrasts, equal variances	C	Yes (NS)	Recommended: planned comparisons or contrasts
LSD	pairwise, equal variances	C	Yes (NS)	Planned pairwise comparisons only
Bonferroni or Dunn-Sidák	pairwise/contrasts, equal variances	E	Yes	Fixed no. unplanned comparisons or contrasts
Scheffé's	pairwise/contrasts, equal variances	E	Yes	Any no. of unplanned comparisons or contrasts or selected contrasts
Tukey's	pairwise only, equal variances	E	Yes	Confidence intervals for pairwise comparisons
Duncan's	pairwise only, equal variances	I	No	Unsuitable for planned or unplanned comparisons
Student-Newman-Keuls (SNK)	pairwise only, equal variances	I	No	Unsuitable for >3 treatments; unplanned pairwise comparisons
Ryan's *Q*	pairwise only, equal variances	E	No	Most powerful test for all pairs comparisons; Recommended
Welsch's Step-Up	pairwise only, equal variances	E	No	Step-up version of Ryan's, slightly less powerful
Waller-Duncan *k*-ratio test	pairwise only, equal variances	I	No	Error rate not always experimentwise, type II error importance hard to set a priori
Kramer's	modification of above tests for unequal *n*	As without modification		Recommended for use with previous tests when *n* unequal
GT2	pairwise only, equal variances, unequal *n*	E	Yes	Less powerful than Tukey/Ryan-Kramer for unequal *n*
T′ method	pairwise only, equal variances, unequal *n*	E	Yes	Less powerful than Tukey/Ryan-Kramer for unequal *n*
Games-Howell (GH)	pairwise only, unequal variances, equal or unequal *n*	E	Yes†	Recommended if unequal variances, error rate can be slightly high
C (modified Cochran's *t*)	pairwise only, unequal variances, equal or unequal *n*	E	Yes	Less powerful than GH, more powerful than T3 if $n < 15$
*T*3	pairwise only, unequal variances, equal or unequal *n*	E	Yes†	Less powerful than GH, more powerful than *C* if $n > 50$
Bryant-Paulson-Tukey (BPT)	pairwise only, on means adjusted in ANCOVA	E	Yes†	Recommended: pairwise comparisons of adjusted means
Conditional Tukey-Kramer	pairwise only, on means adjusted in ANCOVA	E	Yes†	Dependent on covariate values, suitable as for BPT test
Dunnett's	treatments versus control, equal variances	E	Yes‡	Recommended. Use Kramer modification for unequal *n*

† or ‡ This simultaneous test can be used in a stepwise manner for hypothesis tests, with adjusted significance levels († Ryan's; ‡ SNK), to control the experimentwise error rate.

the Fisher's protected LSD test controls the EER at the specified level by virtue of the preliminary overall *F* test. This will only be true when there are no real differences between treatments, as shown by Carmer and Swanson's (1973) own simulation results (see also Keselman et al. 1979, Ryan 1980). Where some means differ, the protected test functions as a simple LSD test for comparisons within groups; it has a per-comparison type I error rate, which is unacceptable for unplanned comparisons.

In the Waller-Duncan *k*-ratio test, the value of the *F* test in the ANOVA is used to set the error rate per comparison (Waller and Duncan 1969; see also Duncan 1965). The critical value is high, with an EER similar to Tukey's test, when the *F* is small; and low, similar to the LSD test with a per-comparison error rate, when the ANOVA *F* is large (Duncan and Brant 1983). This test appears to suffer from the same problem as Fisher's protected LSD test: The EER would not be controlled at the desired level when some of the means differ. The error rate used for a comparison of two treatments would depend on the differences between other treatments. It also seems unlikely that many biologists will have clear-cut grounds for estimating the relative costs of type I and II errors in experiments, as is required for this test. Estimating costs of errors would presumably require that specific alternatives to the overall null hypothesis are considered. If so, then planned comparisons would be more appropriate.

Of those tests that do control the EER, stepwise tests are more powerful than simultaneous tests if significance tests of all pairwise comparisons are required. We have illustrated Ryan's *Q* test as an alternative to Duncan's and the SNK tests because it is the most powerful test that is simple to apply and controls the EER. Ramsey's revised Ryan's test (see Appendix 1) offers a slight gain in power, and Ramsey (1981) has compared the power of this test, used with the *F* rather than the *Q* statistic, with the other new stepwise procedures described in the Introduction. Welsch's Step-

Up test and Shaffer's modification to Welsch's test (used in a step-down manner like Ryan's test) had no power advantage over Ryan's (F) test (Ramsey 1981). We prefer Ryan's test used with Q because it is similar to the methods ecologists are already familiar with. We believe the small gains in power offered in some cases by Peritz's test and the model-testing procedure over Ryan's test (see Einot and Gabriel 1975, Begun and Gabriel 1981, Ramsey 1981) are outweighed by their being much more complex to use.

The simultaneous tests enable simultaneous confidence intervals on the differences between means to be calculated (see below, Presentation and Interpretation of Unplanned Comparisons). These tests are also appropriate when a pair of treatments is compared because the data suggest they are different. As our simulation results show, all the simultaneous UMCPs have actual EERs smaller than the significance level required if the treatment means are not all the same. The most powerful of these UMCPs is Tukey's test. Scheffé's method is always extremely conservative (i.e., it will hold the EER well below the nominal level) because it was not designed for pairwise comparisons only, but for situations where any possible combination of means may be tested (Scheffé 1953). There is an infinite number of such contrasts possible. The Dunn-Sidák method is also always conservative for pairwise comparisons, and therefore also the more commonly used Bonferroni t test, because they rely on inequalities which dictate that the EER must be less than the nominated rate (Neter and Wasserman 1974, Miller 1981).

2. *Pairwise comparisons for unequal sample sizes.*—Dunnett (1980*a*) showed that using the harmonic mean of all the sample sizes, as suggested by Winer (1971) and Snedecor and Cochran (1980), can produce very high EERs and is therefore not recommended. Hayter (1984) has proved that the Kramer modification of Tukey's test for unequal sample sizes will hold the EER at or below the nominal level. Dunnett (1980*a*) and Stoline (1981) show that the Tukey-Kramer procedure produces narrower confidence intervals, i.e., it is more powerful, than the GT2, T', Dunn-Sidák, and Scheffé methods. Gabriel's (1978) approximation to GT2 can result in EERs above the nominal level. Dunnett (1980*a*) presumes that the Kramer modification will also be conservative when applied to the stepwise tests, and our results support this for Ryan's Q test, which remains more powerful than the Tukey-Kramer test when some means differ. Note, however, that no significant difference may be declared if the extreme treatments have small sample sizes, so that the test to compare them has low power. This might prevent detection of differences between closer treatments with larger sample sizes, which would have been significant if tested under the stepwise procedures (Klockers and Sax 1986). The only solution to this "inconsistency problem" is to avoid unequal sample sizes, which would also markedly improve the robustness of all tests to violations of assumptions.

3. *Pairwise comparisons robust to heterogeneous variances (equal or unequal sample sizes).*—Dunnett (1980*b*) has shown that for equal sample sizes the T3 and C procedures are both conservative, even under conditions of extreme variance heterogeneity. $T3$ is more powerful for small sample sizes (roughly $n < 15$), while C is best for $n > 50$. Korhonen (1982), however, has shown that if the sample sizes are very unequal and some are <4, then only C is conservative, but it has very low power. Ecologists generally use small sample sizes (≤ 5), so the $T3$ procedure is probably best, but very uneven sample sizes should be avoided. The GH test can produce EERs slightly above the nominal level (Dunnett 1980*b*), particularly for cases with many treatments (e.g., $m = 8$) and small sample sizes ($n \leq 7$), and small samples with larger variances, but it is often more powerful than the other tests (Keselman and Rogan 1978, Games et al. 1983). Thus GH is recommended where the power of the test is more important than a very strict significance level. The GH and C methods use the Q statistic which appears robust to non-normality (see Assumptions, above). While stepwise versions have not been suggested previously, it seems that Ryan's stepwise significance levels could be applied to these tests. However, the inconsistency problem described for unequal sample sizes will also arise here. Note that in factorial designs these methods cannot be used to compare marginal means as variance pooling assumptions are involved.

4. *Non-pairwise comparisons.*—Scheffé's test controls the experimentwise error correctly for any number of unplanned non-pairwise comparisons. Often the results of an experiment will suggest a choice of one or more comparisons of means. If the choice included, or could have included, comparisons involving combinations of means, then Scheffé's test should be used. However, if the comparison(s) chosen could only have been pairwise, then Tukey's test is appropriate and more powerful.

5. *Non-independence.*—Repeated-measures analyses are still rarely used in ecology and no example of UMCPs following them was found in the literature surveyed. There have been few studies of suitable tests. Jaccard et al. (1984) discussed the tests possible. Maxwell (1980) found that the Bonferroni method remains conservative, whereas the method suggested by Keppel (1973) did not. It should be emphasized that the variance pooling involved in the F tests and comparisons in univariate repeated-measures analyses are very sensitive to the assumptions of homogeneity of variances and covariances, and that multivariate approaches are often more applicable to such designs (see Harris 1985, Gurevitch and Chester 1986).

Twenty-one of the 385 papers that used a posteriori comparisons compared regression slopes or adjusted means after ANCOVA. Normal pairwise UMCPs were used in most cases, although such tests are not appropriate, as we discussed earlier. The use of the GT2 method, as suggested by Sokal and Rohlf (1981), is also

TABLE 9. Summary of suggested or commonly used nonparametric tests for unplanned comparisons of pairs of means. Symbols as in Table 8.

Test	Usage and assumptions	Type I error	Confidence intervals	Recommended/remarks
Mann-Whitney *U* test	pairwise only, equal variances	C	Yes (NS)	Recommended only for planned comparisons
Fligner-Policello	pairwise/contrasts, unequal variances	C	Yes (NS)	Recommended when variances may differ
Steel-Dwass	pairwise only, equal variances	E	Yes	Nonparametric equivalent of Tukey's. Best simultaneous test. Limited exact tables, suspect approximation
Steel-Dwass Ryan	pairwise only, equal variances	E	No	Very low power, due to median ordering
Nemenyi Joint-Rank	pairwise only, equal variances	E	No	Joint ranks, very low power in subgroups; but reasonable tables, easy to use
Joint-Rank Ryan	pairwise only, equal variances	E	No	Joint ranks, but high power, reasonable tables
Dunn's	pairwise only, equal variances	E	No	Approximation to Nemenyi Joint-Rank using unit normal *z* statistic
Steel's	treatments vs. control, equal variances	E	Yes†	Recommended. Fligner modification for unequal *n*

not suitable as it is not robust to unequal variances (Dunnett 1980*a*, *b*, Stoline 1981). Scheffé's test and the Bonferroni methods can be used, but they are generally regarded as too conservative for pairwise comparisons. It is also possible to test the homogeneity of covariances as well as variances (e.g., see Winer 1971): When homogeneity conditions hold for repeated-measures ANOVAs or adjusted means, standard techniques are more powerful than the more robust tests. As for homogeneity of variance tests, however, tests for homogeneity of covariances may not be powerful enough to ensure that the standard techniques are appropriate (see Harris 1985).

We recommend either the Bryant-Paulson-Tukey (BPT) procedure (Bryant and Paulson 1976, Huitema 1980) or the Conditional Tukey-Kramer procedure (Hochberg and Varon-Salomon 1984) for UMCPs on adjusted means. The latter has the advantage of not requiring special tables (the Q statistic is used) and appears to be more powerful than the BPT procedure (Hochberg and Varon-Salomon 1984). When confidence intervals are not required both can be used as stepwise tests using Ryan's adjusted significance levels.

Nonparametric tests (see Table 9).—It is important to note that the usual nonparametric tests in many textbooks assume that the distributions are identical on some scale, except for a difference in medians, in all the treatments compared, so that standard nonparametric ANOVAs and UMCPs should not automatically be used when variances are heterogeneous. The use of the robust Fligner and Policello test when variances are unequal is described in Appendix 1. Also, the null hypothesis being tested by nonparametric UMCPs is not the equality of means but the equality of medians. The relation between means and medians depends on the underlying distribution. Ecologists should be careful not to express the null hypothesis in terms of means when using these tests. Finally, the fundamental difference in rationale between the joint- and pairwise-rank nonparametric UMCPs (see Introduction; Hollander and Wolfe 1973, Miller 1981, Zwick and Marascuilo 1984) should be considered.

As in the parametric case, using two-sample tests (e.g., the multiple Wilcoxon-Mann-Whitney tests used in many ecological papers) will result in high EERs. The test described in Conover (1980) is a nonparametric version of the LSD test and will not control the EER either. "Protecting" these tests with an overall Kruskal-Wallis test will control the EER only when none of the treatments are different (Zwick and Marascuilo 1984), as we and others have shown for the parametric case. Most ecologists did not protect their multiple two-sample tests anyway. Applying stepwise procedures to joint-rank sums (e.g., Zar 1974) will also yield EERs above the nominated level unless re-ranking is carried out for each set of treatments; even then, Duncan's and SNK tests using rank sums will have inflated and indeterminate EERs as in the parametric case (Dodge and Thomas 1980, Campbell and Skillings 1985). However, adjusting the significance levels as in Ryan's Q test will control the EER (our data; Campbell and Skillings 1985).

The simultaneous joint-ranking tests (including those commonly used in the ecological literature, based on the Kruskal-Wallis statistic or Dunn's large-sample approximation), do control the EER, but may have little power except for separating extreme treatments from the rest (our data; Dodge and Thomas 1980, Campbell and Skillings 1985). Simultaneous paired-rank tests (e.g., Steel-Dwass) are more powerful when the sample size is large compared to the number of treatments and especially for two adjacent treatments lying between other treatments (Dunn 1964, Skillings 1983, Fairley and Pearl 1984), a type of difference ecologists often wish to detect. We have shown that the Joint-Rank Ryan test is considerably more powerful than the simultaneous Nemenyi or Steel-Dwass tests (see also Campbell and Skillings 1985) and the Steel-Dwass Ryan test we illustrate. This last test uses the medians to order treatments, which can be inconsistent with the rank sums used for comparisons when treatments are not very different, resulting in very low power.

Although the Steel-Dwass and Steel-Dwass Ryan tests are the nonparametric equivalents of Tukey's and Ryan's tests (using Q or F), they have very limited exact tables and the large sample approximation (Miller 1981, 1986) can be very conservative when there are many treatments (Gabriel and Lachenbruch 1969). Also, these tests can only be used for one-way designs, in contrast to the joint-rank tests and the parametric tests (Miller 1981). A Nemenyi joint-rank test following the Friedman rank two-way analysis is described in Hollander and Wolfe (1973) and Miller (1981).

We therefore recommend that ecologists use the Joint-Rank Ryan test with treatments ordered by their joint-rank sums (this is the ad hoc procedure described by Campbell and Skillings 1985). The all-subset stepwise tests of Campbell and Skillings (1985) may be slightly more powerful but are much more complex to use. As the simultaneous joint-rank tests cannot provide simultaneous confidence intervals, the Steel-Dwass test is the only option when these are required or when a pair of treatments is compared because they look different. It is also more powerful than the simultaneous Nemenyi Joint-Rank test in many cases.

Comparing treatments with a control (see Tables 8 and 9).—For parametric testing of each treatment against a control, Dunnett's test used in a stepwise manner (Miller 1981) is recommended; in this special case (i.e., treatments vs. a control), the SNK significance levels are suitable because groups of means cannot occur. The simultaneous Dunnett's test (described in Winer 1971) is best when confidence intervals are required or treatments are tested because they look different from the control. With unequal sample sizes, use the Kramer modification on Dunnett's test. With unequal variances, apply the GH or $T3$ methods (see above in this section) to Dunnett's test (Dunnett 1985). The method in Dunnett (1964) requires special tables in that paper.

In the nonparametric case, Steel's test (described in Winer 1971, Miller 1981) is suitable, used in a simultaneous or a stepwise manner as in Dunnett's test; Fligner's (1984) modification handles unequal sample sizes. We do not know of a procedure for unequal variances in the nonparametric case, but the Dunn-Sidák method could be applied to the Fligner-Policello test.

Presentation and interpretation of unplanned comparisons

It is most important that ecologists state unambiguously which method was used. Given the variety of strategies possible, authors should explain their choice of methods, particularly in terms of error rates. Using statistical texts or computer packages in a "cookbook" fashion is possibly more dangerous for multiple comparison tests than for most other procedures.

Two approaches can be used when examining pairs of means. One can test a large number of non-orthogonal hypotheses to find which means differ at a given significance level. This is the approach almost invariably used in ecology; we have indicated which UMCPs are best suited for this purpose. When applicable, the stepwise tests are the most powerful. Alternatively, one can admit that there are no specific predetermined hypotheses about how the means may differ, and proceed to describe and interpret the results provided by the experiment using simultaneous confidence intervals. These are intervals that should contain the true differences between the population means. With intervals at the 95% level, for example, there is a probability of 95% that all true differences are contained within the intervals; i.e., the probability refers to the whole collection of intervals, just as the experimentwise error rate refers to the whole collection of significance tests. Huitema (1980) provided a good description of simultaneous confidence intervals (see also Sokal and Rohlf 1981).

Simultaneous confidence intervals are useful ways of presenting and interpreting pairwise comparisons. The estimate of the true difference between means may be an important aspect of the results, or important to readers planning other studies. It is surprising that so much emphasis is placed on the degree of significance of differences between treatments in ecological papers, and so little on estimating the size of the differences. Indeed, in many papers it was impossible to determine the size of the difference(s) from the information presented. If needed, the equivalent of the usual hypothesis test can be done simply by seeing if a difference of zero between two means is contained within the interval, although this will be slightly less powerful than a stepwise hypothesis test such as Ryan's. However, a difference between means other than zero may be biologically important. For example, one may wish to see, using confidence intervals, if a naturally occurring difference is produced by an experimental treatment. Debates on the relative importance of statistical estimation vs. statistical hypothesis-testing are current in much of biology, including entomology (Jones and Matloff 1986, Perry 1986) and medicine (Salsburg 1985, Bailey 1986, Salsburg 1986). Ecologists should keep in mind why they did the experiment(s): Is the size or relative size of differences between treatments important, or is the detection of any differences sufficient? Estimation and significance testing are not mutually exclusive approaches, and could be used together in many analyses.

However the results are analyzed, means, sample sizes, and standard errors should be given so that readers may interpret the results. Such basic information was often not present in the papers we surveyed. This information, coupled with the ANOVA, would alleviate problems such as the use of the Duncan and SNK tests with unknown error rates, since readers could then choose and calculate their own tests.

A common criticism of significance testing in biological papers is that the distinction between biological and statistical significance has been overlooked. Large sample sizes may produce significant statistical results when the differences are trivial in the context of the experiment, and small sample sizes can produce non-significant statistical results in spite of real differences large enough to be important. Prior power analysis would help to avoid the problem, but if simultaneous confidence intervals or sufficient basic information were presented, rather than only the results of tests, readers could assess the importance of differences themselves.

Possibly more than any other type of statistical test, multiple comparisons focus attention on the concepts behind statistical hypothesis testing. In spite of the arguments put forward in some areas of biology and statistics against the usefulness of statistical hypothesis testing, we believe that it has an important role to play in ecology. The current debate brings up three points, however. First, the interpretation of the results of significance tests should be done with more care. While sample sizes in ecology are often small, and thus unlikely to produce significant results that are not biologically important, overreliance on the "religion of significance" (Salsburg 1985) can limit the usefulness of results to other researchers. Second, statistical estimation, particularly of differences between treatments, may be more applicable to ecology than has been appreciated. Finally, since much research in this field is still exploratory, ecologists should consider the techniques of exploratory data analysis, and we refer them to Tukey (1977) and Hoaglin et al. (1983).

Recommendations

1) Enough information should be presented in papers to allow the readers to judge the results for themselves. This is particularly true for potentially ambiguous procedures such as multiple comparisons. At a minimum, means, sample sizes, and standard errors are required.

2) Authors should state unambiguously what tests were used and explain their choice of methods.

3) All assumptions of tests should be considered carefully. Much more emphasis needs to be placed on obtaining equal sample sizes, because nearly all analyses are much more sensitive to assumptions when sample sizes are not equal.

4) Biological knowledge of the situation should be used in deciding whether distributions will be the same across treatments, and whether they are near normal. Plots of residuals or box plots should be used to detect severe non-normality and/or outliers. Use transformations and/or nonparametric tests if non-normality is extreme.

5) Biological data are often lognormal, with unequal variances when the means differ. A plot of means against variances or standard deviations should be used to detect this situation, and indicate whether a log-transform may be appropriate.

6) The assumption of variance equality should be tested if sample sizes are large enough. For normal populations Cochran's test provides reasonable protection, with three treatments, for fairly equal sample sizes ≥ 5. With more treatments slightly larger samples are needed. For unequal sample sizes ≥ 5, an F test of the variance of the smallest sample vs. the average variance of the other samples can be used at the 25% level, in conjunction with Cochran's test. When populations may be non-normal, the Levene-median test is useful for equal sample sizes ≥ 9. Where no variance check is suitable, assume variances are not equal.

7) For unequal variances use transformations if appropriate, or robust ANOVA and comparison methods in situations where these are available. Welch's robust ANOVA, and t test for planned comparisons, are recommended. Standard nonparametric ANOVAs and UMCPs should not be used with unequal variances. In the nonparametric case, use the robust alternatives now available for the Kruskal-Wallis ANOVA and Mann-Whitney U test. Note that the robust methods cannot be used in factorial designs to compare marginal means.

8) Ecologists should always consider whether they can frame specific questions to test, using orthogonal planned comparisons (and associated techniques) with a per-comparison error rate. Non-orthogonal planned comparisons should be kept few in number, and interpreted carefully.

9) When an unplanned comparison is made because the results suggest it, Scheffé's method is appropriate unless the comparison could only have been pairwise, in which case use Tukey's test.

10) For all pairwise comparisons, the parametric Ryan's Q test or the nonparametric Joint-Rank Ryan's test are recommended for hypothesis testing, because they are the most powerful tests which control the experimentwise type I error rate and are also easy to use. Duncan's multiple range test and the SNK test (parametric or nonparametric) are not recommended, because they have an indeterminate error rate. If simultaneous confidence intervals are required, the parametric Tukey's test or the nonparametric Steel-Dwass test are appropriate.

11) When a control is compared to all other treatments, use the parametric Dunnett's test or the nonparametric Steel's test. The more powerful SNK-like stepwise versions of these tests are recommended if simultaneous confidence intervals are not required.

12) For pairwise comparisons with unequal sample sizes, the Kramer modification of Tukey's, Ryan's, and Dunnett's tests is recommended in the parametric case. The Joint-Rank Ryan's test and the approximations to the Steel-Dwass test handle unequal sample sizes, and Fligner's modification applies to Steel's test.

13) For parametric all-pairwise comparisons with unequal variances, the *T*3 procedure seems appropriate for small sample sizes, but it may not be robust with very uneven sample sizes. If a small increase in EER is acceptable, then the Games-Howell test is recommended since it is more powerful, but the user should check if smaller samples have larger variances. For other types of multiple comparisons the parametric Welch (1938) robust *t* test or the nonparametric Fligner-Policello robust Mann-Whitney *U* test can be used with Dunn-Sidák adjusted significance levels.

14) For pairwise comparisons of adjusted means after a parametric ANCOVA, either the Bryant-Paulson generalization of Tukey's or Ryan's *Q* test or the Conditional Tukey-Kramer test are recommended. Standard UMCPs are not suitable. Comparisons after repeated-measures ANOVA should be viewed with caution. The Bonferroni or Dunn-Sidák methods may be appropriate.

Acknowledgments

We would like to thank Neil Andrew, Andrew Brown, Andrew Constable, Peter Fairweather, Prof. John Field, Craig Johnson, Mick Keough, Garry Russ, Prof. Ken Sharpe, Tony Underwood, Ray Watson, Prof. Evan Williams, and particularly Keith McGuinness, for reading drafts of the manuscript and for many helpful discussions. We acknowledge very helpful comments from two unknown referees. This study was supported by the Zoology Department, University of Melbourne and the School of Biological Sciences, University of Sydney.

Literature Cited

Bailey, K. R. 1986. Comment on "the religion of statistics". American Statistician **40**:255–256.

Baker, R. J. 1980. Multiple comparison tests. Canadian Journal of Plant Science **60**:325–327.

Bancroft, T. A. 1968. Topics in intermediate statistical methods. Volume 1. Iowa State University Press, Ames, Iowa, USA.

Bays, C., and S. D. Durham. 1976. Improving a poor random number generator. ACM Transactions on Mathematical Software **2**:59–64.

Begun, J. M., and K. R. Gabriel. 1981. Closure of the Newman-Keuls multiple comparisons procedure. Journal of the American Statistical Assocation **76**:241–245.

Best, D. J., and J. C. W. Rayner. 1987. Welch's approximate solution for the Behrens-Fisher problem. Technometrics **29**:205–210.

Bishop, T. A., and E. J. Dudewicz. 1978. Exact analysis of variance with unequal variances: test procedures and tables. Technometrics **20**:419–430.

Bishop, T. A., and E. J. Dudewicz. 1981. Heteroscedastic ANOVA. Sankhya—the Indian Journal of Statistics **43**:40–57.

Boik, R. J. 1987. The Fisher-Pitman permutation test: a non-robust alternative to the normal theory *F* test when variances are heterogeneous. British Journal of Mathematical and Statistical Psychology **40**:26–42.

Boneau, C. A. 1960. The effects of violations of assumptions underlying the *t* test. Psychological Bulletin **57**:49–64.

Box, G. E. P. 1953. Non-normality and tests on variances. Biometrika **40**:318–335.

———. 1954. Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification. Annals of Mathematical Statistics **25**:290–302.

Box, G. E. P., and M. E. Muller. 1958. A note on the generation of random normal deviates. Annals of Mathematical Statistics **29**:610–611.

Bradley, J. V. 1968. Distribution free statistical tests. Prentice-Hall, Englewood Cliffs, New Jersey, USA.

———. 1977. A common situation conducive to bizarre distribution shapes. American Statistician **31**:147–150.

———. 1978. Robustness? British Journal of Mathematical and Statistical Psychology **31**:144–152.

Brown, R. A. 1974. Robustness of the studentised range statistic. Biometrika **61**:171–175.

Brown, R. A., and A. B. Forsythe. 1974*a*. Robust tests for the equality of variances. Journal of the American Statistical Association **69**:364–367.

Brown, R. A., and A. B. Forsythe. 1974*b*. The small sample behavior of some statistics which test the equality of several means. Technometrics **16**:129–302.

Bryant, J. L., and N. T. Bruvold. 1980. Multiple comparison procedures in the analysis of covariance. Journal of the American Statistical Association **75**:874–880.

Bryant, J. L., and A. S. Paulson. 1976. An extension of Tukey's method of multiple comparisons to experimental designs with random concomitant variables. Biometrika **63**:631–638.

Campbell, G., and J. H. Skillings. 1985. Nonparametric stepwise multiple comparison procedures. Journal of the American Statistical Association **80**:998–1003.

Carmer, S. G., and M. R. Swanson. 1973. An evaluation of ten pairwise multiple comparison procedures by Monte Carlo methods. Journal of the American Statistical Association **68**:66–74.

Carmer, S. G., and W. N. Walker. 1982. Baby bear's dilemma: a statistical tale. Agrononomy Journal **74**:122–124.

Chambers, J. W., W. S. Cleveland, B. Kleiner, and P. A. Tukey. 1983. Graphical methods for data analysis. Duxbury, Boston, Massachusetts, USA.

Chew, V. 1976. Comparing treatment means: a compendium. Hortscience **11**:348–357.

Clinch, J. J., and H. J. Keselman. 1982. Practical alternatives to the analysis of variance. Journal of Educational Statistics **7**:207–214.

Cohen, J. 1977. Statistical power analysis for the behavioral sciences. Academic Press, New York, New York, USA.

Conover, W. J. 1980. Practical nonparametric statistics. Second edition. John Wiley & Sons, New York, New York, USA.

Cox, D. F. 1980. Design and analysis in nutritional and physiological experimentation. Journal of Dairy Science **63**: 313–321.

Damico, J. A., and D. A. Wolfe. 1987. Extended tables of the exact distribution of a rank statistic for all treatments multiple comparisons in one-way layout designs. Communications in Statistics. Part A—Theory and Methods **16**: 2343–2360.

Davenport, J. M., and J. T. Webster. 1975. The Behrens-Fisher problem, an old solution revisited. Metrika **22**:47–54.

Dawkins, H. C. 1983. Multiple comparisons misused: why so frequently in response curve studies? Biometrics **39**:789–790.

Dijkstra, J. B., and P. S. Werter. 1981. Testing the equality of several means when the population variances are unequal. Communications in Statistics. Part B—Simulation and Computation **10**:557–569.

Dodge, Y., and D. R. Thomas. 1980. On the performance of non-parametric and normal theory multiple comparison procedures. Sankhya—the Indian Journal of Statistics **42**: 11–27.

Duncan, D. B. 1955. Multiple range and multiple *F* tests. Biometrics **11**:1–42.

———. 1965. A Bayesian approach to multiple comparisons. Technometrics **7**:171–222.

Duncan, D. B., and L. J. Brant. 1983. Adaptive *t* tests for multiple comparisons. Biometrics **39**:793–794.

Dunn, O. J. 1964. Multiple comparisons using rank sums. Technometrics **6**:241–252.

Dunn, O. J., and V. A. Clark. 1987. Applied statistics: analysis of variance and regression. John Wiley & Sons, New York, New York, USA.

Dunnett, C. W. 1955. A multiple comparisons procedure for comparing several treatments with a control. Journal of the American Statistical Association **50**:1096–1121.

———. 1964. New tables for multiple comparisons with a control. Biometrics **20**:482–491.

———. 1980*a*. Pairwise multiple comparisons in the homogenous variance, unequal sample size case. Journal of the American Statistical Association **75**:789–795.

———. 1980*b*. Pairwise multiple comparisons in the unequal variance case. Journal of the American Statistical Association **75**:796–800.

———. 1982. Robust multiple comparisons. Communications in Statistics. Part A—Theory and Methods **22**:2611–2629.

———. 1985. Multiple comparisons between several treatments and a specified treatment. Pages 39–47 *in* T. Calinski and W. Konecki, editors. Linear statistical inference, lecture notes in statistics. Volume 35. Springer-Verlag, Berlin, Germany.

Einot, I., and K. R. Gabriel. 1975. A study of the powers of several methods of multiple comparisons. Journal of the American Statistical Association **70**:574–583.

Fairley, D., and D. K. Pearl. 1984. The Bahadur efficiency of paired versus joint ranking procedures for pairwise multiple comparisons. Communications in Statistics. Part A—Theory and Methods **13**:1471–1481.

Fisher, R. A. 1935. The design of experiments. First edition. Oliver and Boyd, Edinburgh, Scotland.

Fisher, R. A., and F. Yates. 1953. Statistical tables for biological, agricultural and medical research. Fourth edition. Oliver and Boyd. Edinburgh, Scotland.

Fligner, M. A. 1984. A note on two-sided distribution-free treatment versus control multiple comparisons. Journal of the American Statistical Association **79**:208–211.

Fligner, M. A., and G. E. Policello. 1981. Robust rank procedures for the Behrens-Fisher problem. Journal of the American Statistical Association **76**:162–168.

Gabriel, K. R. 1978. A simple method of multiple comparisons of means. Journal of the American Statistical Association **73**:724–729.

Gabriel, K. R., and P. A. Lachenbruch. 1969. Non-parametric ANOVA in small samples: a Monte Carlo study of the adequacy of the asymptotic approximation. Biometrics **25**:593–596.

Games, P. A., and J. F. Howell. 1976. Pairwise multiple comparison procedures with unequal *n*'s and/or variances: a Monte Carlo study. Journal of Educational Statistics **1**: 113–125.

Games, P. A., J. H. Keselman, and J. C. Rogan. 1983. A review of simultaneous pairwise multiple comparisons. Statistica Neerlandica **37**:53–58.

Games, P. A., and P. A. Lucas. 1966. Power of the analysis of variance of independent groups on non-normal and normally transformed data. Educational and Psychological Measurement **26**:311–327.

Gayen, A. K. 1950. The distribution of the variance ratio on random samples of any size drawn from non-normal universes. Biometrika **37**:236–255.

Gibbons, J. D. 1988. Steel statistics. Pages 751–757 *in* S. Kotz and N. L. Johnson, editors. Encyclopedia of statistical sciences. Volume 8. Wiley-Interscience, New York, New York, USA.

Gill, J. L. 1973. Current status of multiple comparisons of means in designed experiments. Journal of Dairy Science **56**:973–977.

Groggel, D. J., and J. H. Skillings. 1986. Distribution-free tests for main effects in multifactor designs. American Statistician **40**:99–102.

Gurevitch, J., and S. T. Chester. 1986. Analysis of repeated measures experiments. Ecology **67**:251–255.

Harris, R. J. 1985. A primer of multivariate statistics. Academic Press, New York, New York, USA.

Harter, H. L. 1970. Order statistics and their use in testing and estimation. Volume 1. United States Government Printing Office, Washington, D.C., USA.

Hartwig, F., and B. E. Dearing. 1979. Exploratory data analysis. Sage University Paper Series on Quantitative Applications in the Social Sciences. Sage Publications, Beverly Hills, California, USA.

Hays, W. L. 1981. Statistics. Third edition. Holt, Rinehart & Winston, New York, New York, USA.

Hayter, A. J. 1984. A proof of the conjecture that the Tukey-Kramer multiple comparisons procedure is conservative. Annals of Statistics **12**:61–75.

Helwig, J. T., and K. A. Council, editors. 1979. SAS user's guide. 1979 edition. SAS Institute, Raleigh, North Carolina, USA.

Hoaglin, D. C., F. Mosteller, and J. W. Tukey. 1983. Understanding robust and exploratory data analysis. John Wiley & Sons, New York, New York, USA.

Hochberg, Y. 1974. Some generalizations of the *T*-method in simultaneous inference. Journal of Multivariate Analysis **4**:224–234.

Hochberg, Y., and A. C. Tamhane. 1987. Multiple comparison procedures. John Wiley & Sons, New York, New York, USA.

Hochberg, Y., and Y. Varon-Salomon. 1984. Simultaneous pairwise comparisons in analysis of covariance. Journal of the American Statistical Association **79**:863–866.

Hollander, M., and D. A. Wolfe. 1973. Nonparametric statistical methods. John Wiley & Sons, New York, New York, USA.

Huitema, B. E. 1980. The analysis of covariance and its alternatives. Wiley-Interscience, New York, New York, USA.

Hurlbert, R. T., and D. K. Spiegel. 1976. Dependence of *F* ratios sharing a common denominator mean square. American Statistician **30**:74–78.

Jaccard, J., M. A. Becker, and G. Wood. 1984. Pairwise multiple comparison procedures: a review. Psychological Bulletin **96**:589–596.

Jones, D. 1984. Use, misuse, and role of multiple-comparison procedures in ecological and agricultural entomology. Environmental Entomology **13**:635–649.

Jones, D., and N. Matloff. 1986. Statistical hypothesis testing in biology: a contradiction in terms. Journal of Economic Entomology **79**:1156–1160.

Kendall, M. G., and A. Stuart. 1967. The advanced theory of statistics. Hafner, New York, New York, USA.

Kenny, D. A., and C. M. Judd. 1986. Consequences of violating the independence assumption in analysis of variance. Psychological Bulletin **99**:422–431.

Keppel, G. 1973. Design and analysis. A researcher's handbook. First edition. Prentice-Hall, Englewood Cliffs, New Jersey, USA.

———. 1982. Design and analysis. A researcher's handbook. Second edition. Prentice-Hall, Englewood Cliffs, New Jersey, USA.

Keselman, H. J., P. A. Games, and J. C. Rogan. 1979. Protecting the overall rate of Type I errors for pairwise comparisons with an omnibus test statistic. Psychological Bulletin **86**:884–888.

Keselman, H. J., and J. C. Rogan. 1978. A comparison of the Modified-Tukey and Scheffé methods of multiple comparisons for pairwise contrasts. Journal of the American Statistical Association **73**:47–52.

Keuls, M. 1952. The use of the "studentised range" in connection with the analysis of variance. Euphytica **1**:112–122.

Klockers, A. J., and G. Sax. 1986. Multiple comparisons. Sage University Paper Series on Quantitative Applications in the Social Sciences. Sage Publications, Beverly Hills, California, USA.

Knuth, D. E. 1981. The art of computer programming. Second edition. Volume 2. Addison-Wesley, Reading, Massachusetts, USA.

Kohr, R. L., and P. A. Games. 1974. Robustness of the analysis of variance, the Welch procedure and a Box procedure to heterogeneous variances. Journal of Experimental Education **43**:61–69.

Kohr, R. L., and P. A. Games. 1977. Testing complex a priori contrasts on means from independent samples. Journal of Educational Statistics **2**:207–216.

Korhonen, M. P. 1982. On the performance of some multiple comparison procedures with unequal variances. Scandinavian Journal of Statistics **9**:241–247.

Kramer, C. Y. 1956. Extension of multiple range tests to group means with unequal numbers of replications. Biometrics **12**:307–310.

Kurtz, T. E., R. F. Link, J. W. Tukey, and D. L. Wallace. 1965. Short-cut multiple comparisons for balanced single and double classifications: Part 1, Results. Technometrics **7**:95–162.

Levy, K. J. 1978. An empirical comparison of the ANOVA *F* test with alternatives which are more robust against heterogeneity of variance. Journal of Statistical Computation and Simulation **8**:49–57.

Little, T. M. 1978. If Galileo published in Hortscience. Hortscience **13**:504–506.

———. 1981. Interpretation and presentation of results. Hortscience **16**:637–640.

Lund, R. E., and J. R. Lund. 1983. Probabilities and upper quantiles for the studentized range. Applied Statistics **32**: 204–210.

Madden, L. V., J. K. Knoke, and R. Louie. 1982. Considerations for the use of multiple comparison procedures in phytopathological investigations. Phytopathology **72**:1015–1017.

Maxwell, S. E. 1980. Pairwise multiple comparisons in repeated measures designs. Journal of Educational Statistics **5**:269–287.

McGill, R., J. W. Tukey, and W. A. Larsen. 1978. Variations of Box plots. American Statistician **32**:12–16.

Mead, R., and D. J. Pike. 1975. A review of response surface methodology from a biometric viewpoint. Biometrics **31**: 803–851.

Mehta, J. S., and R. Srinivasan. 1970. On the Behrens-Fisher problem. Biometrics **57**:649–655.

Miller, R. G. 1981. Simultaneous statistical inference. Second edition. Springer-Verlag, Berlin, Germany.

———. 1986. Beyond ANOVA, basics of applied statistics. John Wiley & Sons, New York, New York, USA.

Milligan, G. W., D. S. Wong, and P. A. Thompson. 1987. Robustness properties of nonorthogonal analysis of variance. Psychological Bulletin **101**:464–470.

Morley, C. L. 1982. A simulation study of the powers of three multiple comparison statistics. Australian Journal of Statistics **24**:201–210.

Nemenyi, P. 1963. Distribution-free multiple comparisons. Dissertation. Princeton University, Princeton, New Jersey, USA. Cited in Hollander and Wolfe 1973.

Neter, J., and W. Wasserman. 1974. Applied linear statistical models. Richard D. Irwin, Homewood, Illinois, USA.

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

O'Neill, R. T., and B. G. Wetherill. 1971. The present state of multiple comparisons methods. Journal of the Royal Statistical Society, Series B **33**:218–241.

Patel, K. M., and D. G. Hoel. 1973. A nonparametric test for interaction in factorial experiments. Journal of the American Statistical Association **68**:615–620.

Perry, J. N. 1986. Multiple-comparison procedures: a dissenting view. Journal of Economic Entomology **79**:1149–1155.

Peterson, R. G. 1977. Use and misuse of multiple comparison procedures. Agronomy Journal **69**:205–208.

Petronovich, L. F., and C. D. Hardyck. 1969. Error rates for multiple comparison methods: some evidence concerning the frequency of erroneous conclusions. Psychological Bulletin **71**:43–54.

Posten, H. O. 1984. Robustness of the two-sample *t* test. Pages 92–99 *in* D. Rasch and M. L. Tiku, editors. Robustness of statistical methods and nonparametric statistics. D. Reidel, Dordrecht, The Netherlands.

Ramsey, P. H. 1978. Power differences between pairwise multiple comparisons. Journal of the American Statistical Association **73**:479–485.

———. 1981. Power of univariate pairwise multiple comparison procedures. Psychological Bulletin **90**:352–366.

Ramseyer, G. C., and T. Tcheng. 1973. The robustness of the studentised range statistic to violations of the normality and homogeneity of variance assumptions. American Educational Research Journal **10**:235–240.

Ratcliffe, J. F. 1968. The effect on the *t* distribution of nonnormality in the sampled population. Journal of Applied Statistics **17**:42–48.

Renner, B. R., and D. W. Ball. 1983. The effects of unequal variances on the Tukey WSD test. Educational and Psychological Measurement **43**:27–34.

Rivest, L-P. 1986. Bartlett's, Cochran's and Hartley's tests on variances are liberal when the underlying distribution is long-tailed. Journal of the American Statistical Association **81**:124–128.

Robson, D. S. 1959. A simple method for constructing orthogonal polynomials when the independent variable is unequally spaced. Biometrics **15**:186–191.

Rogan, J. C., and H. J. Keselman. 1977. Is the ANOVA *F*-test robust to variance heterogeneity when sample sizes are equal? An investigation via a coefficient of variation. American Educational Research Journal **14**:493–498.

Rogan, J. C., H. J. Keselman, and L. J. Breen. 1977. Assumption violations and rates of type 1 error for the Tukey multiple comparison test: a review and empirical investigation via a coefficient of variance variation. Journal of Experimental Education **46**:20–26.

Rohlf, F. J., and R. R. Sokal. 1981. Statistical tables. Second edition. W. H. Freeman, San Francisco, California, USA.

Rosen, M. R., and B. F. Hoffman. 1978. Statistics, biomedical scientists, and circulation research. Circulation Research **42**:739.

Rosenthal, R., and R. L. Rosnow. 1985. Contrast analysis: focused comparisons in the analysis of variance. Cambridge University Press, Cambridge, England.

Rust, S. W., and M. A. Fligner. 1984. A modification of the Kruskal-Wallis statistic for the generalized Behrens-Fisher problem. Communications in Statistics. Part A—Theory and Methods **13**:2013–2027.

Ryan, T. A. 1959. Multiple comparisons in psychological research. Psychological Bulletin **56**:26–47.

———. 1960. Significance tests for multiple comparisons of proportions, variances, and other statistics. Psychological Bulletin **57**:318–328.

———. 1980. Comment on "Protecting the overall rate of Type I errors for pairwise comparisons with an omnibus test statistic". Psychological Bulletin **88**:354–355.

Salsburg, D. S. 1985. The religion of statistics as practiced in medical journals. American Statistician **39**:220–257.

———. 1986. Reply. American Statistician **40**:256–257.

Satterthwaite, F. E. 1946. An approximate distribution of estimates of variance components. Biometrics Bulletin **2**: 110–114.

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

———. 1959. The analysis of variance. John Wiley & Sons, New York, New York, USA.

Shaffer, J. P. 1977. Multiple comparisons emphasizing selected contrasts: an extension and generalization of Dunnett's procedure. Biometrics **33**:293–303.

———. 1979. Comparison of means: an *F* test followed by a modified multiple range procedure. Journal of Educational Statistics **4**:14–23.

Skillings, J. H. 1983. Nonparametric approaches to testing and multiple comparisons in a one-way ANOVA. Communications in Statistics. Part B—Simulation and Computation **12**:373–387.

Snedecor, G. W., and W. G. Cochran. 1980. Statistical methods. Seventh edition. Iowa State University Press, Ames, Iowa, USA.

Sokal, R. R., and F. J. Rohlf. 1969. Biometry. First edition. W. H. Freeman, San Francisco, California, USA.

Sokal, R. R., and F. J. Rohlf. 1981. Biometry. Second edition. W. H. Freeman, San Francisco, California, USA.

Spjøtvoll, E., and M. R. Stoline. 1973. An extension of the *T*-method of multiple comparison to include cases with unequal sample sizes. Journal of the American Statistical Association **68**:975–978.

Steel, R. G. D. 1959. A multiple comparison rank sum test: treatment versus control. Biometrics **15**:560–572.

———. 1960. A rank sum test for comparing all pairs of treatments. Technometrics **2**:197–207.

———. 1961. Some rank sum multiple comparisons tests. Biometrics **17**:539–552.

Steel, R. G. D., and J. H. Torrie. 1960. Principles and procedures of statistics. McGraw-Hill, New York, New York, USA.

Still, A. W., and A. P. White. 1981. The approximate randomization test as an alternative to the *F* test in analysis of variance. British Journal of Mathematical and Statistical Psychology **34**:243–252.

Stoline, M. R. 1981. The status of multiple comparisons: simultaneous estimation of all pairwise comparisons in one-way ANOVA designs. American Statistician **35**:134–141.

Tamhane, A. C. 1977. Multiple comparisons in model 1 one-way ANOVA with unequal variances. Communications in Statistics Part A—Theory and Methods **6**:15–32.

———. 1979. A comparison of procedures for multiple comparisons of means with unequal variances. Journal of the American Statistical Association **74**:471–480.

Tan, W. Y., and M. A. Tabatabai. 1986. Some Monte-Carlo studies on the comparison of several means under heteroscedasticity and robustness with respect to departure from normality. Biometrical Journal **28**:801–814.

Thigpen, C. C., and A. S. Paulson. 1974. A multiple range test for analysis of covariance. Biometrika **61**:479–484.

Thomas, D. A. H. 1973. Multiple comparisons among means, a review. The Statistician **22**:16–42.

Tomarken, A. J., and R. C. Serlin. 1986. Comparison of ANOVA alternatives under variance heterogeneity and specific non-centrality structures. Psychological Bulletin **99**: 90–99.

Tukey, J. W. 1953. Some selected quick and easy methods of statistical analysis. Transactions of the New York Academy of Sciences, Series 2, **16**:88–97.

———. 1977. Exploratory data analysis. Addison-Wesley, Reading, Massachusetts, USA.

Underwood, A. J. 1981. Techniques of analysis of variance in experimental marine biology and ecology. Oceanography and Marine Biology Annual Review **19**:513–605.

Ury, H. K. 1976. A comparison of four procedures for multiple comparisons among means—pairwise contrasts for arbitrary sample sizes. Technometrics **18**:89–97.

Waldo, D. R. 1976. An evaluation of multiple comparison procedures. Journal of Animal Science **42**:539–544.

Waller, R. A., and D. B. Duncan. 1969. A Bayes rule for the symmetric multiple comparisons problem. Journal of the American Statistical Association **64**:1484–1503.

Wang, Y. Y. 1971. Probabilities of the type 1 errors of the Welch tests for the Behrens-Fisher problem. Journal of the American Statistical Association **66**:605–608.

Warren, W. G. 1979. Response: Analysis and interpretation of an experiment with a heterogeneous mixture of treatment types. Biometrics **35**:869–872.

Welch, B. L. 1938. The significance of the difference between two means when the population variances are unequal. Biometrika **29**:350–362.

———. 1947. The generalization of 'Students' problem when several different population variances are involved. Biometrika **34**:28–35.

———. 1951. On the comparison of several mean values: an alternative approach. Biometrika **38**:330–336.

Welsch, R. E. 1977. Stepwise multiple comparison procedures. Journal of the American Statistical Association **72**: 566–575.

Wilcox, R. R. 1986. Controlling power in a heteroscedastic ANOVA procedure. British Journal of Mathematical and Statistical Psychology **39**:65–68.

———. 1987. Pairwise comparisons of *J* independent regression lines over a finite interval, simultaneous pairwise comparison of their parameters, and the Johnson-Neyman procedure. British Journal of Mathematical and Statistical Psychology **40**:80–93.

Wilcox, R. R., V. L. Carlin, and K. L. Thompson. 1986. New Monte-Carlo results on the robustness of the ANOVA *F*, *W* and *F** statistics. Communications in Statistics. Part B—Simulation and Computation **15**:933–943.

Williams, D. A. 1972. The comparison of several dose levels with a zero dose control. Biometrics **28**:519–531.

Winer, B. J. 1971. Statistical principles in experimental design. Second edition. McGraw-Hill, New York, New York, USA.

Zar, J. H. 1974. Biostatistical analysis. First edition. Prentice-Hall, Englewood Cliffs, New Jersey, USA.

———. 1984. Biostatistical analysis. Second edition. Prentice-Hall, Englewood Cliffs, New Jersey, USA.

Zwick, R., and L. A. Marascuilo. 1984. Selection of pairwise multiple comparison procedures for parametric and nonparametric analysis of variance methods. Psychological Bulletin **95**:148–155.

APPENDIX 1

This section provides the necessary formulae to implement the tests that are commonly used or that we recommend. Parametric tests compare treatment means. In nonparametric tests, the data in all treatments in the experiment (joint ranking), or the data in the two treatments in each pairwise comparison (pairwise ranking), are combined into one set and ranked from smallest to largest. These tests compare treatment mean ranks or rank sums. An alternative approach to pairwise nonparametric tests, based on the Wilcoxon-Mann-Whitney test, is to use placements (the position of an observation in one sample relative to a second sample).

The following general notation will be used in these formulae:

a is the chosen (i.e., nominal) significance level; e.g., .05 or 5%.

b is an adjusted significance level used to keep the experimentwise level at a.

CV is the critical value for each test. The difference between treatments must be greater than this value to be declared significant.

df is the degrees of freedom, and for parametric tests this is the df of the standard error (SE$_c$). df* is Satterthwaite's adjusted degrees of freedom (see Welch's t test, below in this Appendix).

r is the number of comparisons to be made.

m is the number of means (or treatments) in the experiment.

n_i is the number of replicates in treatment i.

Q is the Studentized Range statistic, t is Student's t statistic, F is the F (variance ratio) statistic; these are tabulated in most standard texts.

For parametric tests:

SE$_c$ = the standard error of the comparison being tested. For pairwise comparisons with equal n and homogeneous variances, SE$_c = \sqrt{(2\text{MS}_d/n)}$, where MS$_d$ is the mean square in the denominator of the ANOVA F ratio. For heterogeneous variances use SE$_c$* (see Welch's t test, below in this Appendix). For other cases, see below.

For non-parametric tests:

R_i = rank of an observation in treatment i.

$\sum R_i$ = sum of ranks in treatment i.

$\bar{R}_i = \sum R_i/n_i$; i.e., mean rank of treatment i.

Consider a pair of treatments (i and j), and let $n_i > n_j$. Then:

P_i = the "placement" of an observation in treatment i. Then P_i is the number of observations in treatment j less than this observation in treatment i.

$\sum P_i$ = the sum of placements for sample i. Rank sums can be converted to placement sums by $\sum P_i = \sum R_i - n_i(n_i + 1)/2$.

$\bar{P}_i$ = mean placement for treatment i ($= \sum P_i/n_i$).

For pairwise rankings of treatments i and j, convert maximum to minimum rank or placement sums using: $\sum R_i = (n_i + n_j)(n_i + n_j + 1)/2 - \sum R_j$ and $\sum P_i = n_i n_j - \sum P_j$.

1. Parametric Planned Comparisons

Specify the comparison as a linear combination of means $L_{\bar{X}} = c_1\bar{X}_1 + \ldots c_i\bar{X}_i + \ldots c_m\bar{X}_m$, where $\bar{X}_i$ is the mean and c_i the coefficient of treatment i. $L_{\bar{X}}$ is a valid comparison (with one degree of freedom) if $\sum c_i$ (or $\sum n_i c_i$ for unequal n_i) = 0. Two comparisons, L_a and L_b, are orthogonal if $\sum c_{ai}c_{bi}$ (or $\sum n_i c_{ai} c_{bi}$ for unequal n_i) = 0.

Equal variances (sample sizes equal or unequal)

In the formulae below, SE$_c = \sqrt{[\text{MS}_d \sum (c_i^2/n_i)]}$, which reduces to $\sqrt{(2\text{MS}_d/n)}$ for pairwise comparisons with equal n. Either use means, where CV for the comparison $= t_{a(\text{df})}\text{SE}_c$, or calculate a sum of squares (SS) with 1 df, and hence also a mean square, in the context of the ANOVA: SS $= L_T^2/\sum n_i c_i^2$, where L_T is a comparison of treatment totals (with means replaced by totals in the linear combination). Test this against the appropriate MS$_d$ in an F test.

Unequal variances (sample sizes equal or unequal)

Welch's (1938) Approximate t test (using Satterthwaite's adjusted df):

$CV = t_{a(\text{df*})}\text{SE}_c$*, where SE$_c$* $= \sqrt{\sum c_i^2 s_i^2/n_i}$ and df* $= (\sum c_i^2 s_i^2/n_i)^2/\sum \{c_i^4 s_i^4/[n_i^2(n_i - 1)]\}$. Standard t tables are used. For pairwise comparisons these formulae become SE$_c$* $= \sqrt{(s_i^2/n_i + s_j^2/n_j)}$, and df* $= (s_i^2/n_i + s_j^2/n_j)^2/\{s_i^4/[n_i^2(n_i - 1)] + s_j^4/[n_j^2(n_j - 1)]\}$.

2. Nonparametric Planned Comparisons

Equal variances (sample sizes equal or unequal)

Mann-Whitney-Wilcoxon test (comparing treatments i and j, treatment i is largest):

Calculate $U = \sum R_i - n_i(n_i + 1)/2$, or $U = \sum P_i$, and compare $U^* =$ maximum of U and $n_i n_j - U$ with tabulated values in most standard texts.

Unequal variances (sample sizes equal or unequal)

Fligner-Policello test (comparing treatments i and j, where treatment i is largest):

Calculate

$$U = (\sum P_i - \sum P_j)/$$
$$[2\sqrt{\sum (P_i - \bar{P}_i)^2 + \sum (P_j - \bar{P}_j)^2 + \bar{P}_i\bar{P}_j}]$$

and use tables in Fligner and Policello (1981).

3. Parametric Unplanned Comparisons—All Pairs

Equal sample sizes and variances

Simultaneous tests.—Note that, for most tests (see Table 7) simultaneous confidence intervals on differences between all pairs of treatment means can be calculated by $\bar{X}_i - \bar{X}_j \pm CV$, at the chosen significance level a.

LSD (Least Significant Difference) test:

$CV = t_{a(\text{df})}\text{SE}_c$ or $\sqrt{F_{a(1,\,\text{df})}}\text{SE}_c$ or $Q_{a(2,\,\text{df})}\text{SE}_c/\sqrt{2}$. In pairwise multiple t tests the same formulae apply, except that SE$_c$ is calculated from the standard errors of the means in each pair.

Bonferroni method:

$CV = t_{b(\text{df})}\text{SE}_c$, where the adjusted significance level $b = a/r$.

Dunn-Sidák method:

CV as in Bonferroni method except $b = 1 - (1 - a)^{1/r}$.

Scheffé's test:

$CV = \sqrt{[(m - 1)F_{a(m-1,\,\text{df})}]}\text{SE}_c$.

Tukey's test:

$CV = Q_{a(m,\,\text{df})}\text{SE}_c/\sqrt{2}$.

Stepwise tests.—For all the tests below, $CV = Q_{b(p,\,\text{df})}\text{SE}_c/\sqrt{2}$, where p is the number of means in the group to be tested and b is the adjusted significance level for a test of the equality of p means.

For Duncan's (new) multiple range test, $b = 1 - (1 - a)^{p-1}$.
For the Student-Newman-Keuls (SNK) test, $b = a$.
For Welsch's step-up test (GAPA), $b = ap/m$ except $b = a$ when $p = m - 1$.
For Ryan's Q test, $b = 1 - (1 - a)^{p/m}$. This is the test proposed by Einot and Gabriel (1975); and because $p \leq m$, $ap/m \leq 1 - (1 - a)^{p/m}$, so that the test using $1 - (1 - a)^{p/m}$ is more powerful.

For Ramsey's revised Ryan's test, $b = 1 - (1 - a)^{p/m}$ for $p < m - 1$ and $b = a$ for $p = m, m - 1$.

Note: To obtain table values of Q_b where $b = 1 - (1 - a)^{p/m}$ one must interpolate in the available tables. When the 5% significance level is chosen ($a = .05$) and ≤ 10 means are compared, calculate $Q_b = Q_{.05} + (Q_{.01} - Q_{.05})[\ln(.05) - \ln(b)]/[\ln(.05) - \ln(.01)]$ where "ln" denotes natural logarithms. For more means or other nominal significance levels, similar interpolations can be based on Q tables in Harter (1970) or Zar (1974). Alternatively, Q values between 1 and 5% may be calculated from an algorithm in Lund and Lund (1983). It is easy to program the interpolation to produce a set of "Ryan's Q tables" which can be used in the same way as the Q tables.

Unequal sample sizes and equal variances

The Kramer Modification of Tukey's test and stepwise tests uses the harmonic mean of the two sample sizes, $2/(1/n_i + 1/n_j)$, so that $\mathrm{SE}_c = \sqrt{[\mathrm{MS}_d(1/n_i + 1/n_j)]}$.

Unequal variances (equal or unequal sample sizes)

Games and Howell (GH) method:

$CV = Q_{a(m,\ \mathrm{df}^*)}\mathrm{SE}_c{}^*/\sqrt{2}$.

T3 method:

$CV = \mathrm{SMM}_{a(m,\ \mathrm{df}^*)}\mathrm{SE}_c{}^*$.

These tests use $\mathrm{SE}_c{}^*$ and Satterthwaite's adjusted degrees of freedom, df*, as for Welch's (1938) approximate t test (above, Parametric Planned Comparisons: Unequal Variances). SMM is the Studentized Maximum Modulus statistic, tabulated in Rohlf and Sokal (1981).

C method:

$CV = [Q_{a(m,\ \mathrm{df})}s_i^2/n_i + Q_{a(m,\ \mathrm{df})}s_j^2/n_j]/\sqrt{[2(s_i^2/n_i + s_j^2/n_j)]}$

Non-independence in one-way ANCOVA

Let $S = \sqrt{\mathrm{MS}_{\mathrm{res}}[1 + (\mathrm{MS}_{\mathrm{g_x}}/\mathrm{SS}_{\mathrm{r_x}})]}$, where $\mathrm{MS}_{\mathrm{res}}$ is the residual mean square from the ANCOVA, $\mathrm{MS}_{\mathrm{g_x}}$ is the between-groups (or treatments) mean square from the ANOVA on the covariate (x), $\mathrm{SS}_{\mathrm{r_x}}$ is the residual sum of squares from the ANOVA on the covariate (x).

Bryant-Paulson-Tukey test for adjusted means:

$CV = Qp_{a(m,\ \mathrm{df})}S/\sqrt{n}$, where Qp is the generalized studentized range distribution, tabulated in Huitema (1980).

Conditional Tukey-Kramer test (Hochberg and Varon-Salomon 1984):

$CV = Q_{a(m,\ \mathrm{df})}\sqrt{\mathrm{MS}_{\mathrm{res}}}\sqrt{[1/n + (\bar{X}_i - \bar{X}_j)^2/2\mathrm{SS}_{\mathrm{r_x}}]}$, where $\bar{X}_i$ and $\bar{X}_j$ are the two covariate means of the treatments for which adjusted means are being compared.

Note: Both tests can be used in stepwise manner, substituting Qp_b for Qp_a and Q_b for Q_a where $b = 1 - (1 - a)^{p/m}$ as in Ryan's Q test.

4. Nonparametric Unplanned Comparisons— All Pairs, Joint Ranking

Equal or unequal sample sizes and equal variances

Simultaneous.—Nemenyi Joint-Rank test:

Use the absolute value of differences between mean ranks ($|\bar{R}_i - \bar{R}_j|$) of treatments in the joint ranking of all treatments. For $m \leq 4$ and various combinations of n between 2 and 6, compare $N^*(|\bar{R}_i - \bar{R}_j|)$ with exact CVs tabulated in Damico and Wolfe (1987), where N^* is the lowest common multiple of sample sizes. Further exact CVs for $|\bar{R}_i - \bar{R}_j|$ for small (and equal) n with $m \leq 11$ are in Hollander and Wolfe (1973: Table A9); multiplication by N^* is unnecessary for these tables. For other combinations of m and n, compare $|\bar{R}_i - \bar{R}_j|$ with the large-sample approximation $Q_{a(k,\ \mathrm{df}=\infty)}\sqrt{[\sum n(\sum n + 1)/12]}\sqrt{[1/2(1/n_i + 1/n_j)]}$ where $\sum n$ is the total number of observations in all treatments.

Stepwise.—Joint-Rank Ryan test:

Use test as above with adjusted significance levels, as in Ryan's Q test, where $b = 1 - (1 - a)^{p/m}$. Note that at each p-mean level, the new subset of treatments must be reranked.

Unequal variances

UMCPs may be calculated from the robust Kruskal-Wallis procedure recently described by Rust and Fligner (1984), but the formulae are complex.

5. Non-parametric Unplanned Comparisons— All Pairs, Pairwise Ranking

Equal sample sizes and variances

Simultaneous—Steel-Dwass test:

Compare each pair of treatments, ranked separately, with Mann-Whitney-Wilcoxon tests (see Nonparametric Planned Comparisons, above). For $m = 3$ and $n = 2$ to 6, convert calculated U to minimum rank sum (see general notation) and compare with exact CVs in Steel (1960). For larger m and/or n, convert calculated U to maximum rank sum and compare with the large sample approximation $n(2n + 1)/2 + 1/2 + Q_{a(m,\ \mathrm{df}=\infty)}\sqrt{[n^2(2n + 1)/24]}$; some values are tabulated in Steel (1961) and Miller (1981: Table IX).

Stepwise.—Steel-Dwass Ryan test:

Use the Steel-Dwass test with adjusted significance levels, as in Ryan's Q test, where $b = 1 - (1 - a)^{p/m}$. Treatments should be ordered by medians.

Unequal sample sizes and equal variances

Exact tables are not available for the Steel-Dwass test with unequal sample sizes. Calculate $U^* = U/n_in_j$ and compare with large sample approximation (Miller 1986, based on Dunn 1964): $(1/2) + Q_{m,\ \mathrm{df}=\infty)}\sqrt{[(n_i + n_j + 1)/(24n_in_j)]}$

Unequal variances

Use pairwise Fligner-Policello tests (see Nonparametric Planned Comparisons, above) and adjust significance levels with Dunn-Sidák method.

Simultaneous confidence intervals (Gibbons 1988)

There are n_in_j differences between each observation in treatment i and each observation in treatment j. Arrange these differences (D_t) in ascending order so that $D_t{}^Y$ is the difference ranked "Y." Calculate $L = 1 + [n_j(2n_i + n_j + 1)]/2 - r_t{}^*$, where $r_t{}^*$ is the CV of the Steel-Dwass test (maximum rank sum). Calculate $H = n_in_j + 1 - L$. Lower confidence limit $= D_t{}^L$ (i.e., the difference ranked "L" in ascending order). Upper confidence interval $= D_t{}^H$ (i.e., the difference ranked "H" in ascending order). Repeat for each pair of treatments.

6. Treatments vs. A Control—Parametric

In the formulae below, n_0 is sample size of control and n_t is sample size of treatment t.

Equal sample sizes and variances

Dunnet's test:

Simultaneous.—For each treatment vs. control pair, $CV = d_{(m,\ \mathrm{df})}\mathrm{SE}_c$ where m (number of treatments) includes the control and d is found in tables for Dunnett's test (Dunnett 1964, Winer 1971, Zar 1974, 1984).

Stepwise.—Use Dunnett's test by comparing treatment furthest from control first, then next furthest from control, etc., as in an SNK test (SNK significance levels are suitable for this case).

Unequal sample sizes and equal variances

Use Dunnett's test with the Kramer modification: $\mathrm{SE}_c = \sqrt{\mathrm{MS}_d(1/n_0 + 1/n_t)}$

Unequal variances

Apply GH or C methods (see above) to Dunnett's test.

Simultaneous confidence intervals (each treatment mean vs. control mean)

Upper and lower confidence limits $= \bar{X}_t - \bar{X}_0 \pm d_{(m,\,\mathrm{df})}\mathrm{SE}_c$, where SE_c as for unequal n_0 and n_t.

7. Treatments vs. a Control—Nonparametric

Equal treatment sample sizes (can be different from control sample size) and equal variances

Steel's test with Fligner's (1984) modification:

Simultaneous.—For each treatment vs. control, find max of $\sum R_t$ and $n_t(n_0 + n_t + 1) - \sum R_t$ and compare with $CV = n_t(n_0 + n_t + 1)/2 + 1/2 + d_{(m-1,\,\infty)}\sqrt{[n_t n_0(n_t + n_0 + 1)/12]}$, where $d_{(m-1,\,\infty)}$ is from tables for Dunnett's test (Dunnett 1964, Winer 1971, Zar 1974, 1984), with df $= \infty$ and $(m - 1)$ means (including control). Steel (1959) provides a few exact values of the minimum rank sum. Approximate tables in Steel (1959) and Miller (1981) assume $n_0 = n_t$ and are conservative, as they are calculated from earlier tables for Dunnett's test.

Stepwise.—Use Steel's test as for stepwise Dunnett's test.

Unequal treatment sample sizes and equal variances

Fligner (1984) provides a solution based on iterative solving of a complex formula for the distribution. An approximate test for small sample sizes is to replace $d_{(m-1)}$ in the formula above by the Mann-Whitney U at a significance level of $a/2(m - 1)$.

Unequal variance

No test described. Use Fligner-Policello test (see Nonparametric Planned Comparisons, above) and adjust significance levels with the Dunn-Sidák method for $m - 1$ comparisons.

Simultaneous confidence intervals (Fligner 1984, Gibbons 1988)

As for Steel-Dwass except that there are $n_0 n_t$ differences between each observation in treatment t and each observation in the control, and $L = 1 + [n_t(2n_0 + n_t + 1)/2] - r_t^*$, where r_t^* is CV of Steel's test (maximum rank sum) for treatment t vs. control. Repeat for each treatment vs. control.

8. Alternatives to the Parametric ANOVA

Welch ANOVA (robust to unequal variances; Welch 1951):

$W = (m + 1) \sum [w_i(A\bar{x}_i - \sum w_i\bar{x}_i)^2]/[A^2(m^2 - 1) + 2(m - 2)B]$, where $w_i = n_i/s_i^2$; $A = \sum w_i$; $B = \sum [(A - w_i)^2/(n_i - 1)]$; degrees of freedom $= m - 1$, $A^2(m^2 - 1)/3B$. Compare with F table.

APPENDIX 2

Fictitious data are presented to illustrate points about the tests used. An experiment examining the effects of four surface types on the recruitment of an intertidal animal (e.g., barnacles) was set up (see Underwood 1981 for a similar example) (Table A1).

We illustrate parametric and nonparametric analyses of the data, each with three alternative strategies to compare treatments: (a) planned comparisons, (b) stepwise unplanned comparisons of all treatment pairs, and (c) simultaneous unplanned comparisons of pairs (with simultaneous confidence intervals on the differences). The planned comparisons are tests of three specific hypotheses relevant to the experiment: (1) the numbers of recruits are the same on the two types of algae, (2) the numbers of recruits are the same on natural bare rock and scraped areas, and (3) the numbers of recruits are the same on algae and bare areas. Unplanned comparisons would only be carried out in the absence of such specific hypotheses. Confidence intervals would be used if estimates of the differences were required. Formulae for the tests are in Appendix 1.

TABLE A1. Barnacle data. Two treatments were types of algae (A1 and A2) that occurred naturally on the rocks; one treatment was naturally bare rock (NB); the last was a rock surface which had been manually scraped clean (S), to see how closely this technique represented natural bare rock. There were five replicate areas of each of the four treatments, and after 4 wk the number of barnacles that had recruited to each area was recorded.

	Treatments			
	A1	A2	NB	S
	27	24	9	12
	19	33	13	8
	18	27	17	15
	23	26	14	20
	25	32	22	11
Mean	22.4	28.4	15.0	13.2
Variance	14.8	15.3	23.5	20.7
Total	112	142	75	66

1. Parametric Analyses

Checks of assumptions

Although the small sample size ($n = 5$) does not permit powerful tests of normality or homogeneity of variances, rough preliminary checks should still be carried out. Plots of residuals, the absence of any positive relationship between means and variances, and Cochran's test ($C = 0.32$, $P > .05$) suggested there was no serious violation of normality or variance homogeneity. ANOVA is thus appropriate (Table A2).

Planned comparisons

Totals are used to partition the treatments sum of squares (SS) into three orthogonal (see below) comparisons; each comparison SS is tested against residual mean square (MS) using an F test (1,16 df). Alternatively, means and t tests could be used as shown in comparison number 1.

1) H_0: there is no difference in the number of barnacles between the two algal surfaces (treatments A1 and A2). $L_T = (+1)142 + (-1)112$, where $+1$ and -1 are coefficients (c_i). Note that $\sum c_i = (+1) + (-1) = 0$, so this is a legitimate comparison with 1 df; SS $= 90.0$, $F = 4.84$, df $= 1{,}16$, $P < .05$.

or

TABLE A2. ANOVA.

Source of variation	SS	df	MS	F ratio	P
Treatments	736.55	3	245.52	13.22	<.05
Residual	297.20	16	18.58		
Total	1033.75	19			

L(of means) = (+1)28.4 + (−1)22.4, where +1 and −1 are coefficients. $L = 6.00$, SE = 2.726, $t = 2.20$, df = 16, $P < .05$. Note that $F = t^2$.

2) H_0: there is no difference in the number of barnacles between the naturally bare and scraped surfaces (treatments NB and S). $L_i = (+1)75 + (-1)66$; ss = 8.1, $F = 0.44$, df = 1,16, NS. Because $\sum c_{1i}c_{2i} = 0$ [(+1)(0) + (−1)(0) + (0)(+1) + (0)(+1) = 0], H_{02} is orthogonal to H_{01}.

3) H_0: there is no difference in the number of barnacles between algal-covered surfaces and bare surfaces (i.e., between the average of treatments A1 and A2 and the average of treatments NB and S). $L_T = (+1/2)142 + (+1/2)112 + (-1/2)75 + (-1/2)66$; ss = 638.5, $F = 34.45$, $P < .001$. H_{03} is orthogonal to H_{01} and H_{02} because $\sum c_{1i}c_{3i} = 0$ and $\sum c_{2i}c_{3i} = 0$.

Note: ss(H_{01}) + ss(H_{02}) + ss(H_{03}) = treatments ss, and there are three comparisons with 1 df, and three treatments df. This is a check for orthogonal comparisons.

Stepwise unplanned comparisons

Overall question: Are there differences between pairs of treatments?

Ryan's Q test

4 means apart: $Q_b = 4.05$, $CV = 7.81$, A2 − S = 15.2, $P < .05$.
3 means apart: $Q_b = 3.85$, $CV = 7.42$, A1 − S = 9.2, $P < .05$, A2 − NB = 13.4, $P < .05$.
2 means apart: $Q_b = 3.48$, $CV = 6.71$, NB − S = 1.8, $P > .05$, A1 − NB = 7.4, $P < .05$, A2 − A1 = 6.0, $P > .05$.
Result: S NB < A1 A2 (lines join treatments not significantly different, $P > .05$).

Simultaneous unplanned comparisons

Overall question: Are there differences between pairs of treatments?

Tukey's test

$Q_a = 4.04$, $CV = 7.79$
A2 − S = 15.2, $P < .05$, A2 − NB = 13.4, $P < .05$,
A2 − A1 = 6.0, $P > .05$, A1 − S = 9.2, $P < .05$,
A1 − NB = 7.4, $P > .05$, NB − S = 1.8, $P > .05$.

Result: S NB < A1 A2.

95% Simultaneous confidence intervals

The probability is at least 95% that the true differences between all pairs of population means are within the following ranges:

A2 − A1 = −1.79 to 13.79 (i.e., 6.0 ± 7.79); A1 − NB = −0.39 to 15.19; NB − S = −5.99 to 9.59; A2 − NB = 5.61 to 21.19; A1 − S = 1.41 to 16.99; A2 − S = 7.41 to 22.99.

2. Nonparametric Analyses

Nonparametric analyses are used when the approximate normality of the data cannot be confirmed. Nonparametric methods test for differences between population medians. Note that checks of the variance homogeneity assumption are important for this analysis. Plots of residuals are suitable, but since the data are presumably not normal, Cochran's test would not be used. For larger sample sizes, the Levene-median test would be appropriate.

Summary of the data: the treatment medians are: S = 12.0, NB = 14.0, A1 = 23.0, and A2 = 27.0.

Kruskal-Wallis test, using Chi-square approximation = 13.614 (tie-corrected = 13.625), $P < .01$.

Planned comparisons

Planned comparisons are handled the same as for the parametric case except for the use of Wilcoxon-Mann-Whitney tests, here using placements. In cases 1 and 2, $n_1 = n_2 = 5$; in case 3, $n_1 = n_2 = 10$.

1) A2 vs. A1, $U = 21.5$, NS.
2) NB vs. S, $U = 16$, NS.
3) A2 + A1 vs. NB + S, $U = 96$, $P < .001$.

Stepwise unplanned comparisons

Joint-Rank Ryan test (difference between mean ranks), using Miller's (1981, 1986) large sample approximation. New joint ranks must be determined for each set in the stepwise procedure. Treatments should be ordered according to joint rank sums.

Sets of 4 treatments: $Q_b = 3.63$, $CV = 9.62$; A2 − S = 12.1, $P < .05$.
Sets of 3 treatments: $Q_b = 3.45$, $CV = 9.14$; A1 − S = 7.0, $P > .05$, A2 − NB = 8.9, $P > .05$.
Sets of 2 treatments: no tests.

Result: S NB A1 A2.

Steel-Dwass Ryan test (maximum rank sum of each pair of treatments), using Miller's (1981) large sample approximation (nearest integer). Treatments should be ordered according to sample medians.

4 medians apart: $Q_b = 3.63$, $CV = 40$; A2 − S = 40, $P < .05$
3 medians apart: $Q_b = 3.45$, $CV = 40$; A1 − S = 38, $P > .05$, A2 − NB = 40, $P < .05$
2 medians apart: $Q_b = 3.14$, $CV = 39$; A2 − A1 = 35, $P > .05$

Result: S NB A1 A2.

Simultaneous unplanned comparisons

Nemenyi Joint-Rank test, using Miller's (1981, 1986) large sample approximation. Joint ranks in 4 treatment set used for all comparisons.

$Q_a = 3.63$, $CV = 9.60$ (difference between mean ranks).
A2 − S = 12.1, $P < .05$; A1 − S = 7.7, $P > .05$;
A2 − NB = 10.7, $P < .05$; NB − S = 1.4, $P > .05$;
A1 − NB = 6.3, $P > .05$; A2 − A1 = 4.4, $P > .05$.
Result: S NB A1 A2.

Steel-Dwass test, using Miller's (1981) large-sample approximation.

$Q_a = 3.63$, $CV = 40$ (maximum rank sum).
A2 − S = 40, $P < .05$; A1 − S = 38, $P > .05$;
A2 − NB = 40, $P < .05$, NB − S = 31, $P > .05$;
A1 − NB = 38, $P > .05$; A2 − A1 = 35, $P > .05$.
Result: S NB A1 A2.

95% Simultaneous confidence intervals
(pairwise rankings must be used)

These are confidence intervals on differences between medians. $CV = 40$ (maximum rank sum); $L = 25$, $H = 1$; the confidence limits are the 1st- and 25th-ranked differences between observations in any pair of treatments. Therefore the probability is at least 95% that the true differences between all pairs of population medians are within the following ranges:

S − NB = −11 to 14; A1 − S = −2 to 19; A2 − S = 4 to 25; A1 − NB = −4 to 18; A2 − NB = 2 to 24; A2 − A1 = −3 to 14.

3. Welch ANOVA for Unequal Variance Case

This is calculated in stages below, following the formula in Appendix 1.

w_1 = .242, w_2 = .213, w_3 = .338, w_4 = .327;
$\bar{x}_1$ = 13.2, $\bar{x}_2$ = 15.0, $\bar{x}_3$ = 22.4, $\bar{x}_4$ = 28.4;
A = 1.119, m = 4, B = .707;
W = 11.837, df$_1$ = 3, df$_2$ = 8.8 (approximately = 9), CV (F table) = 3.86; Significant, $P < .05$.