Nonparametric Analysis of Ordinal Data in Designed Factorial Experiments

D. A. Shah and L. V. Madden

Department of Plant Pathology, New York State Agricultural Experiment Station, Cornell University, Geneva, NY 14456; and Department of Plant Pathology, Ohio State University, Wooster 44691.
Accepted for publication 26 August 2003.

ABSTRACT

Shah, D. A., and Madden, L. V. 2004. Nonparametric analysis of ordinal data in designed factorial experiments. Phytopathology 94:33-43.

Plant disease severity often is assessed using an ordinal rating scale rather than a continuous scale of measurement. Although such data usually should be analyzed with nonparametric methods, and not with the typical parametric techniques (such as analysis of variance), limitations in the statistical methodology available had meant that experimental designs generally could not be more complicated than a one-way layout. Very recent advancements in the theoretical formulation of hypotheses and associated test statistics within a nonparametric framework, together with development of software for implementing the methods, have made it possible for plant pathologists to analyze properly ordinal data from more complicated designs using nonparametric techniques. In this paper, we illustrate the nonparametric analysis of ordinal data obtained from two-way factorial designs, including a repeated measures design, and show how to quantify the effects of experimental factors on ratings through estimated relative marginal effects.

Additional keywords: distribution-free methods, normalized distribution, rank-based methods.

Before beginning an experiment, researchers must decide on the experimental design, the measurement scale to be used in quantifying the response of individuals to imposed treatments, and ideally, the method of analyzing and expressing the obtained data. These decisions are interconnected, because the appropriate statistical methods for analyzing the data depend on the measurement scale chosen and experimental design (59,60). In fact, the chosen design may restrict which analyses are possible, and the measurement scale can determine which designs should be used. Continuous measurement scales (sometimes called interval or ratio scales [23]) are popular among plant pathologists, partly because this type of scale gives data that can be analyzed fairly easily by so-called parametric statistical methods (e.g., analysis of variance [ANOVA] and *t* tests) with few or no imposed limitations on the type of experimental design that can be used.

Nevertheless, plant pathologists face many situations in which the use of a continuous scale of measurement is time consuming or impractical. Assessing severity of disease for many of the root pathogens certainly falls into this category, as does the assessment of virus disease severity in many plants (20,46,49). In these and similar situations, researchers have often used an ordinal scale of measurement (10,21,25,29,42,48,54,56,67). An ordinal scale is one in which the values used for the measurement are interpretable only in terms of their arrangement in a given order, for example, from least to most severe. Therefore, a 0-to-4 scale, in which 0 = no disease, 1 = slight, 2 = moderate, 3 = severe, and 4 = completely killed, only has meaning within the conceptual interpretation that the disease severity levels have an intrinsic order to them. One could just as well use the labels 0 = no disease, 10 = slight, 50, = moderate, 100 = severe, and 120 = completely killed. Many disease assessment keys are ordinal (27,34,57,61).

It is easy to see that, with ordinal scales, differences between the measured values are not interpretable, at least in a quantitative sense. Regardless of the value labels chosen, the difference between slight and moderate is not necessarily the same as the difference between severe and completely killed. Furthermore, means (and differences between means) based on these value labels cannot be interpreted in the same sense as the means of observations measured on a continuous scale. Parametric methods of analysis using statistics based on means, or differences between means (such as ANOVA), are thus, strictly speaking, inappropriate for analyzing data on an ordinal scale, though they are used quite often in many disciplines (47).

The statistical approach for dealing with ordinal data should be invariant to the choice of transformation or measurement scale used, that is, it should maintain the original order of the values on the ordinal scale (2). Analysis based on rank transformations meets these criteria. Given an observation (x) in a set of observations, the rank of x is the number of observations that are less than or equal to x. Ranks represent the underlying order of the values and are invariant under monotonic transformations. That is, if one used values from the 0-to-4 scale or 0-to-120 scale, or the square root of the values, the ranking of the observations would not change. Differences in ranks are more readily interpretable (if the difference in the ranks of two observations is 3, then there are exactly two observations between them) than the differences between ordinal values. The Kruskal-Wallis test (23,37) is a popular rank-based statistical method of analysis and is the nonparametric equivalent of the one-way ANOVA. The Friedman test (15,23) is a rank-based method for a randomized complete block design with one experimental factor (plus a blocking factor).

Although classical rank-based methods have been researched and used over a fairly long time, they were generally not satisfactory beyond the one-way layout for experiments. That is, they could not be used for factorials, split plots, nested and repeated

Corresponding author: L. V. Madden; E-mail address: madden.1@osu.edu

*The *e*-Xtra logo stands for "electronic extra" and indicates that the online article contains supplemental material not included in the print edition. The online supplement contains detailed instructions on use of statistical software and interpretation of the output, SAS programs (with data), and annotated output files.

Publication no. P-2003-1113-01O

measures designs (2,3,16). Moreover, some nonparametric methods developed for higher-order layouts were not purely rank-based (7) and were not invariant under monotonic transformations. An apparent exception was the rank transform method, in which the original observations are replaced by their ranks and the usual parametric ANOVA performed on the ranks themselves (31). Several criticisms of the rank transform method have been published based on theoretical, analytical, and methodological arguments (2,7,8,16,58,64,65). First, hypotheses tested in ANOVA are based on differences between means (i.e., decomposition of means [or mean ranks]), or shifts in the means, which are affected by monotonic transformations of the data. However, rank statistics are invariant to monotonic transformation, so it is inappropriate to use rank statistics to test hypotheses that depend on the transformation (3). Second, ranked data will not be normally distributed, violating the standard assumption of normality for the popular parametric tests. Third, ranked data generally will have unequal variances (2), even if the original data have constant variance, a further and serious violation of the assumption of typical parametric tests.

One simple example is given here of the problems that can occur in testing certain hypotheses of interest with the rank transform method. Consider an experiment with two crossed factors, A and B, in which both factors affect the expected (mean) response (i.e., both main effects are significant), but that the interaction of A and B has no effect on the expected response (e.g., that the effect of B on the expected value does not depend on the level of A). A type I error occurs if a test for interaction is significant in this scenario. Blair et al. (8) showed with detailed simulations with normally distributed data that there can be a very high type I error rate in testing for interactions using the rank transform method, and the error rate increases as the magnitude of the main effects increases. Furthermore, the error rate increases as sample size increases. The type I error rate can well exceed 50%, and even approach 100% under many realistic circumstances, indicating that the test results for interaction are essentially worthless in this type of situation (58).

Research on rank-based methods eventually led to the realization that the problem lay in the proper definition of treatment or factor effects. With continuous scale data and ANOVA (or *t* tests), effects are straightforward and unambiguous, being defined with expected (mean) values (estimated by arithmetic averages or by the ANOVA model being used). For a one-way layout, for instance, the null hypothesis being tested is that expected value (μ) is the same for all treatments. The effects of the treatments are quantified by the differences in estimated μ values. This formulation for treatment effects, unfortunately, does not carry over to the nonparametric situation for factorials and other complicated experimental designs. However, once a suitable definition of what is now called nonparametric effects and hypotheses was established in a highly significant paper by Akritas and Arnold (4), a unified approach was paved to the analysis of all ordinal data, as well as other data that did not meet the assumptions of parametric statistical analysis. Much progress was made in the late 1990s on the unified theory for rank statistics by E. Brunner and colleagues, extending methods of analysis for ordinal data beyond the one-way layout to factorials, split plots, and repeated measures (i.e., data collected over time on experimental units) (5,6,11,13,14,18). Many of the advancements are summarized in two recent books (12,15) and a recent review (17).

The nonparametric methodology of Brunner and colleagues represents a significant advancement in the statistical analysis of ordinal data, because as we have pointed out, the approach is generalizable for many experimental designs. Now that statistical software is available for performing the calculations, researchers outside of statistics can conduct appropriate analysis and interpret their results, provided they are taught how to conduct the analysis. In this paper, we demonstrate the nonparametric analysis of ordinal data from designed experiments using the new rank-based methods. We first present the basic concepts behind the analysis, using minimal mathematics, and then demonstrate the methodology by analyzing published data sets. We first consider a one-way layout as a relatively simple situation to introduce and explain several relevant concepts. Then, we consider a two-way (crossed) layout as well as a repeated measures design and discuss how to apply the method to a split-plot design. Finally, we place the nonparametric analysis presented in this article in a broader statistical framework for the analysis of ordinal data. We hope this presentation will show that researchers now have alternatives to ANOVA and other parametric methods for analyzing ordinal disease rating data.

ONE-WAY LAYOUT: CONCEPTS AND EXAMPLE

Nonparametric effects and hypotheses. In the nonparametric framework, there are no parameters (such as the mean or variance) from distributions (e.g., normal) on which to base treatment effects and hypotheses. One approach that is generalizable to many situations is to define treatment effects in reference to the distributions of the variables measured in the experiment (4,5). Suppose we have an experiment with a treatments (i = 1, …, a) and n_i independent experimental units or replications (k = 1, …, n_i) per treatment. We place an i subscript on n because each treatment could have a different number of replications. The measurement in the kth replication of treatment i is represented by X_{ik}, which is a random variable with normalized distribution $F_i(x)$. For instance, the measurement in the third replicate of treatment 2 is X_{23}. Lower case x is used to indicate any specific value of the random variable X. For data without ties, the distribution (sometimes called the cumulative distribution) is the probability that X is less than or equal to a specific value, x. To handle data with or without ties, the normalized distribution, $F_i(x)$, is used instead of the simpler distribution function (36). The formula for $F_i(x)$ is given by Brunner et al. (page 46 of literature citation 12). The normalized distribution can represent any type of random variable, including continuous and ordinal categories. It is assumed (for now) that the X_{ik} values are independent.

The weighted average of all the $F_i(x)$s in an experiment, $H(x)$, is given by

$$H(x) = \frac{1}{N}\sum_{i=1}^{a} n_i F_i(x) \quad (1)$$

where N is the total number of observations ($= \sum_1^a n_i$). Each distribution is multiplied by the number of replications in equation 1, so that more weight is given to groups (treatments) with more information (i.e., independent observations) than groups with less. When the number of replications is the same for each treatment, then $H(x)$ is a simple average of the normalized distributions. For ease of presentation, we generally abbreviate $H(x)$ and $F_i(x)$ as H and F_i, respectively. Using H and F_i, we define the so-called relative effect for the ith treatment as

$$p_i = \int H dF_i \quad (2)$$

in which dF_i is the first derivative of F_i. The p_i value describes the so-called stochastic tendency of F_i with respect to H (17,18). Specifically, if p_i is >1/2, observations in the ith treatment tend to be larger in comparison to an independent random variable that has H as its distribution; likewise, if p_i is <1/2, observations in the ith treatment tend to be smaller. A p_i of 1/2 indicates that there is no tendency for observations from the ith treatment to be larger or smaller in comparison to a random variable with H as its distribution. Differences in the p_i values are used to compare treatments.

For continuous data, dF_i is just the probability density function for the ith treatment. With a normal distribution, dF_i represents the classic bell-shaped Gaussian curve taught in introductory statistics

classes. It may be interesting to note that if x is substituted for H in equation 2, the expected value (i.e., mean [μ_i]) for the ith treatment is obtained. In one sense, therefore, one can think of the relative treatment effect as a generalized expectation or mean.

The relative treatment effects, p_i, are estimated by replacing the distribution functions H and F_i by their corresponding empirical distributions, $\hat{H}$ and $\hat{F}_i$. The direct calculations are quite tedious for this (pages 45 to 52 in Brunner et al. [12]), although users of the methodology do not need to be concerned with this step. It can be shown that the estimated relative treatment effect ($\hat{p}_i$) can be determined directly from the observation midranks (with midranks, if there are three tied values, for example, they would all have the same rank). For brevity, we generally refer to midranks simply as ranks. If R_{ik} is the rank of X_{ik} among all N observations, then the mean rank for the ith treatment can be written as

$$\overline{R}_{i\bullet} = \frac{1}{n_i}\sum_{k=1}^{n_i} R_{ik} \tag{3}$$

The dot subscript indicates that the average over all the replications for the ith treatment is calculated. The relative treatment effect can then be estimated from the mean rank as

$$\hat{p}_i = \frac{1}{N}\left(\overline{R}_{i\bullet} - \frac{1}{2}\right) \tag{4}$$

Thus, there is a straightforward link between the mean ranks used in many nonparametric analyses and the relative effects of this method. Note that the ranks are not of direct interest here for characterizing observations from individual treatments, but are used "as a natural and convenient tool for estimating the relative treatment effects" (12). The $\hat{p}_i$ has an asymptotic normal distribution (6), but its variance or standard error [$se(\hat{p}_i)$] when the null hypothesis is not true (i.e., when the relative effects are not all equal) is a very complicated expression that generally requires specialized software to calculate (17).

To demonstrate some of these concepts, consider the data in Figure 1, which corresponds to a simple situation of two treatments (i.e., a = 2) with 15 observations per treatment (i.e., n_1 = n_2 = 15; k = 1, …, 15). The observations (X_{1k} and X_{2k} for treatments 1 and 2, respectively) are shown in the inset box of the figure. The normalized distributions were estimated from these data using the methods described by Brunner et al. (pages 45 to 47 in literature citation 12) and shown on the graph. Although the methods in this paper are based on the normalized distributions, one usually does not need to directly estimate and graph these for data analysis purposes; however, their presentation can be useful for understanding the concepts behind the analysis. The standard distribution and normalized distribution are the same here due to the lack of ties in this demonstration data set. Because the values of X were generally smaller in treatment 1 than in treatment 2, the estimated distribution for the first treatment ($\hat{F}_1$) is to the left of the one for the second treatment ($\hat{F}_2$). The estimated weighted average distribution ($\hat{H}$) at each observed value (at each of the 2·15 = 30 observations) is estimated directly using the ranks of the observations: $\hat{H}(X_{ik}) = (R_{ik} - 1/2)/N$ (page 50 in Brunner et al. [12]). Mean ranks (based on equation 3) were $\overline{R}_{1\bullet}$ = 9.5 and $\overline{R}_{2\bullet}$ = 21.5. Estimated medians were 3.0 and 5.7 for treatments 1 and 2, respectively. Moreover, using equation 4, estimated relative treatment effects were $\hat{p}_1$ = 0.3 and $\hat{p}_2$ = 0.7. Thus, based on these p_i estimates, two equivalent statements can be made: (i) observations from treatment 2 tend to take on larger values in comparison to an independent random variable that has H as its distribution; or (ii) observations from treatment 2 tend to take on larger values compared with observations from both treatments combined. This is analogous to the normal-theory-based situation, in which the expected (mean) value of treatment 2 is larger than the overall mean of the two treatments ($\overline{\mu}$). We emphasize here that the estimates of the relative effects are not affected by monotonic transformations of the observations. For example, taking the square root or the logarithm of the data in Figure 1 does not change the estimates of $\hat{p}_1$ (=0.3) and $\hat{p}_2$ (=0.7).

For the special case of a one-way layout with only two treatments, interpretation of the relative effects can be even more direct. Instead of considering the effects relative to a mean weighted distribution (H), one can consider the relative effect of F_2 with respect to F_1 (with $p = p_2 - p_1 + 1/2$), essentially ignoring H. Here, p is the probability that the response variable from treatment 2 is greater than the response variable from treatment 1. Equivalently, when p is >1/2, observations from treatment 2 tend to be larger (are "tendentiously larger") than observations from treatment 1 (12). Graphically, F_2 tends to be to the right of F_1 (Fig. 1) when p is >1/2. If there are no differences in the distributions for the two treatments, $p_2 = p_1 = 1/2$, which means that p equals 1/2. For the example data in Figure 1, $\hat{p} = 0.7 - 0.3 + 0.5 = 0.9$, indicating a strong tendency for observations in treatment 2 to be larger than those in the first treatment and for the F_1 curve to be to the left of the F_2 curve. This is supported by the estimates of the distribution functions ($\hat{F}_1$, $\hat{F}_2$).

Hypotheses and tests. Statistical hypotheses for treatment effects are formulated in terms of the normalized distributions. The simplest null hypothesis (H_0) is that the treatments have no effect at all. This can be written as the null hypothesis (H_0^F)

$$H_0^F : F_1 = F_2 = \ldots = F_a \tag{5}$$

Using the example in Figure 1, the null hypothesis is simply that $F_1 = F_2$, that is, that the two distributions are the same (overlap completely). The data that were used to construct the estimated distributions ($\hat{F}_1$, $\hat{F}_2$) are utilized to test this hypothesis (described below). The alternative hypothesis is that at least one of the relative treatment effects (p_i) is different from the rest. Thus, when treatments do not have equal distributions, the treatment differences are quantified by differences in relative effects. Readers experienced in ANOVA will recognize that equation 5 is simply a generalization of the usual null hypothesis of a linear-model

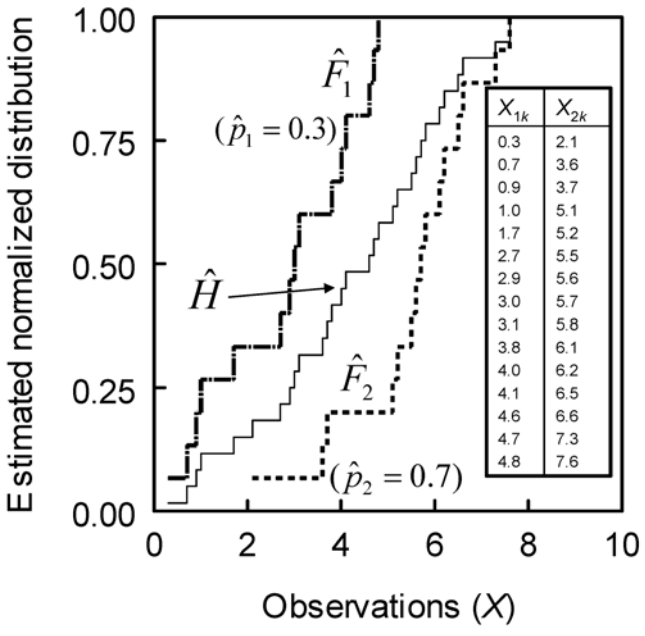

Fig. 1. Estimated (normalized) distribution function, also known as the empirical distribution function, for two treatments (labeled 1 and 2; thick broken lines) and the estimated weighted distribution function (thin solid line). The lines are "step-like" in appearance because the estimates were determined just for the observed data points, shown in the insert. Data were obtained with a random number generator. Estimated relative treatment effects ($\hat{p}_i$) are shown in the frame.

analysis for normally distributed data that is discussed in introductory statistics books:

$$H_0^{\mu} : \mu_1 = \mu_2 = \ldots = \mu_a \quad (6)$$

where μ_i is the mean (expected value) for the ith treatment. The null hypothesis of equation 5, however, does not depend on any distributional parameters and is applicable for any measurement scale of the data.

Analysis of data is based on the observed ranks, R_{ik}. Because the rank is a nonlinear transformation of the data, variances of the ranks will generally vary with treatment (2,17). This is a major reason why the standard rank transform method of Conover (23) and Conover and Iman (24) is generally inappropriate for analyzing ordinal data (16), as discussed in the introduction. In the approach of Brunner et al. (13) and Brunner and Puri (17,18), the rank transformation is not regarded as a technique for the derivation of statistics (as it is with the Conover and Iman method [24]), but as a property of a statistic that can be useful for computational purposes (12). Two types of statistics can be used for testing the null hypothesis shown in equation 5 (17). The first is the so-called Wald-type statistic (WTS), which has, asymptotically, a chi-square distribution (with $a - 1$ degrees of freedom [df = $a - 1$]) under the null hypothesis. However, very large sample sizes are needed to achieve a good approximation, and the test does not perform well with small or moderate sample sizes (6,12,17). The second statistic is known as the so-called ANOVA-type statistic (ATS), which has, based on asymptotic theory, an approximate F distribution (with df_N [numerator] and df_D [denominator] degrees of freedom) under the null hypothesis. The approximation—developed by Brunner et al. (11) based on the ideas of Box (9)—holds for all but very small sample sizes. Degrees of freedom for this statistic are determined from complicated expressions based on the number of treatment levels, number of observations, and the variance of ranks in each treatment (18). The df_D calculation is, essentially, an extension of the method used for a two-sample t test when the variances of the two samples are not equal, a method taught in introductory statistics courses. Often, df_N is somewhat less than $a - 1$, and df_D is somewhat less than the corresponding value for a one-way ANOVA F test on the original data.

In addition to the tests of the general null hypothesis, other contrasts can be tested with WTS and ATS. For instance, the difference in normalized distributions between two treatments or two groups of treatments (e.g., 1 and 2 versus 3 and 4) can be evaluated in terms of their relative treatment effects. An additional test statistic, called the linear rank statistic (L) has also been derived for testing so-called patterned alternative hypotheses (12). For instance, if treatments correspond to increasing inoculum dose applied to plants, one can test the alternative hypothesis that there is a linear increase in the relative treatment effects of the disease rating distributions (relative to the null hypothesis of equation 5). The test statistic for patterned alternatives has a standard normal distribution under the null hypothesis (equation 5), which is approximated by a t distribution at small sample sizes.

Software. The estimation of nonparametric treatment effects and the tests of hypotheses for the approach outlined above require the calculation of midranks and the ability to specify heteroscedastic variance structures in the model. These can be done easily with SAS (SAS Institute, Cary, NC), using PROC RANK to first obtain the midranks and then PROC MIXED, with appropriate options, to specify the heteroscedastic variance model and request the proper test statistics. PROC MIXED is a very general procedure for fitting linear mixed models to data, but can be used for many specific purposes once the modeling conventions are understood.

Additionally, Brunner and colleagues have written SAS macros to perform the calculations for several different experimental designs. The macros are downloadable for free online (available on the website of E. Brunner at the University of Göttingen, Germany). PROC MIXED cannot be used to calculate standard errors for the estimates of the relative treatment effects. However, a separate macro can be used to estimate the standard errors. Recently, the SAS macros have been converted to run on the R statistical program (The R Foundation for Statistical Computing, Vienna, Austria) for those who do not have access to SAS.

Detailed instructions on the use of PROC MIXED and macros for the nonparametric analysis, and advice on the interpretation of output, are available in electronic format from both authors.

One-way layout example. We illustrate here how to use SAS to analyze data nonparametrically for the one-way layout. Later, we show how to analyze data from more complicated designs.

Verticillium dahliae Kleb. is a soilborne pathogen that infects potato (*Solanum tuberosum* L.), causing potato early dying (PED). Omer et al. (52) compared several different isolates of *V. dahliae* based on the severity of PED on potato cv. Superior. Their study was done in the greenhouse and was a completely randomized design, with isolate as the treatment factor. Severity of foliar symptoms was recorded on a 1-to-6 ordinal scale: 1 = no visible symptoms, 2 = slight chlorosis of lower leaves, 3 = extensive chlorosis of lower leaves, 4 = extensive chlorosis and some necrosis of lower and upper leaves, 5 = severe stunting with chlorosis and necrosis of entire plants, and 6 = dead or nearly dead plants. We examine here a subset of their data from test 1, consisting of six *V. dahliae* isolates from Montana. Isolates were either of the 4A or 4B vegetative compatibility groups. There were eight replications of each isolate.

The data set therefore consists of the following variables: isolate (six levels), rating (ordinal rating score between 1 and 6), and subject (i.e., a unique identifier for each experimental unit; this variable is needed when calculating confidence intervals). Although data were collected over 6 weeks, we use the data only from week three.

Isolate had a significant effect on the distribution of rating values based on the test statistics. The WTS was 135.4 (df = 5), which was significant at $P < 0.001$. The ATS was 10.10 (df_N = 3.86 and df_D = 30.6), which was also significant at $P < 0.001$. The results for the WTS probably should not be used because of the moderate sample sizes (eight replications) here, but are shown for comparison. The median disease ratings per isolate ranged from 1.0 to 3.5, which resulted in $\hat{p}_i$ values ranging from 0.23 to 0.88 (Table 1). The analysis is based on all the data and not medians, but the median does provide one convenient (and traditional) summary of the central value for each treatment. The meaning of the $\hat{p}_i$ values can be put in perspective by considering two of the isolates. The value of 0.88 (corresponding to the largest median rating) indicates that observations for the first isolate tended to be larger than those represented by the weighted mean distribution H. The value of 0.23 (corresponding to the smallest median rating) indicates that observations for the fifth isolate tended to be smaller than observations with H for a distribution. Thus, the first and fifth isolates were the farthest apart in terms of estimated relative treatment effects. The confidence intervals were calculated based on the (estimated) standard error (se) according to the method of Brunner et al. (page 60 of literature citation 12). The approach is slightly more complicated than simply adding or subtracting $z \cdot \mathrm{se}(\hat{p}_i)$, where z is the standard normal variable at the specified confidence level, to $\hat{p}_i$. Corrections are made to ensure that the limits of the interval fall between the maximum and minimum possible p_i values $[n_i/(2N), 1 - n_i/(2N)]$. For this data set, the range is [8/96, 1–8/96] or [0.083, 0.917]. One can see in Table 1, among other things, that there is no overlap of the estimated relative treatment effects for the first and fifth isolates. The contrast of the two vegetative compatibility groups resulted in an ATS of 20.47 (df_N = 1 and df_D = 30.6; $P < 0.001$), indicating that there was a significant difference in disease between

the two groups. In addition to predetermined contrasts, pairwise comparison of relative treatment effects can also be done using *t* tests of $\hat{p}_i$ differences, in which the standard error of the difference of two estimated relative treatment effects is determined as the square root of the sum of the squares of the two individual standard errors.

The analysis of the PED data set also could have been done with the Kruskal-Wallis test because there was only one experimental factor. In fact, it should be noted that for the one-way layout, the Kruskal-Wallis test is a special case of the more general relative treatment effects test (17), in which the variance of ranks (under the null hypothesis) is the same for all levels of the treatment, $\sigma_0^2 = N(N + 1)/12$. Analysis of data from a one-way layout serves as a convenient way of introducing the general methodology without the complications of multiple factors.

FACTORIAL EXPERIMENTAL DESIGNS

Concepts. One can directly generalize the relative treatment effect analysis for a single experimental factor to two or more factors. The concepts are explained here for the situation with two factors, which we label A and B. We suppose that the experiment has *a* levels of factor A ($i = 1, \ldots, a$) and *b* levels of factor B ($j = 1, \ldots, b$). There are n_{ij} experimental units or replications for each combination of the levels of A and B, and the index *k* is used to refer to a specific replicate ($k = 1, \ldots, n_{ij}$). The subscript *ij* is placed on *n* because the number of replications can vary with the levels of the experimental factor. For example, if there are five field-plot replications corresponding to the first level of A ($i = 1$) and third level of B ($j = 2$), then n_{13} equals 5. If $a = 4$ levels of factor A and $b = 3$ levels of factor B, then there is a total of $a \times b = 12$ combinations of factor levels (i.e., there are 12 unique *ij* combinations). The measurement in the *k*th replication of the *ij* factor level combination is represented by X_{ijk}, which is a random variable with normalized distribution F_{ij}. Unlike the situation with normal data, however, no assumptions are made about the form of F_{ij}. The weighted average of all the distributions (i.e., for the total of $a \times b$ combinations of the two factors) is given by

$$H = \frac{1}{N}\sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}F_{ij} \tag{7}$$

As with the one-way layout, more weight is given to distributions with more information (i.e., more observations) than to distributions with less information. In a similar way, one can define the mean (normalized) distribution for the *i*th level of A across the levels of B ($\overline{F}_{i\bullet}$) and the mean normalized distribution for the *j*th level of B across the levels of A ($\overline{F}_{\bullet j}$). These are theoretical values and are analogous to the main-effect expected (mean) values (e.g., $\mu_{i\bullet}$ and $\mu_{\bullet j}$) that one considers in an ANOVA for normal data. Direct estimates of F_{ij}, $\overline{F}_{i\bullet}$, and $\overline{F}_{\bullet j}$ can be obtained (pages 45 to 47 in Brunner et al. [12]) but are not generally needed for the analysis.

The relative effect of the *i*th level of A and *j*th level of B (i.e., the *ij* combination of the two factor levels) is given by

$$p_{ij} = \int H dF_{ij} \tag{8}$$

The relative effect, p_{ij}, known as the relative marginal effect in factorials, measures the difference of the distribution for the *ij*th combination (F_{ij}) from the weighted mean distribution. These relative effects can be estimated very simply from the observation midranks without directly using the estimates of *H* and F_{ij} (12). If R_{ijk} is the rank of X_{ijk}, then one can write the mean rank for a specific combination of levels of A and B (e.g., for *ij*) as

$$\overline{R}_{ij\bullet} = \frac{1}{n_{ij}}\sum_{k=1}^{n_{ij}} R_{ijk} \tag{9}$$

As with equation 3, the dot subscript indicates an average over the n_{ij} replications. The relative effect is then estimated as

$$\hat{p}_{ij} = \frac{1}{N}\left(\overline{R}_{ij\bullet} - \frac{1}{2}\right) \tag{10}$$

One can also estimate *p* for the *i*th level of factor A across all the levels of B and for the *j*th level of factor B across all the levels of A using "main effect" mean ranks [$\hat{p}_{i\bullet} = (1/N)(\overline{R}_{i\bullet\bullet} - 1/2)$; $\hat{p}_{\bullet j} = (1/N)(\overline{R}_{\bullet j\bullet} - 1/2)$]. Calculation of the standard error of the estimated relative marginal effect, $se(\hat{p}_{ij})$, is quite complicated when the null hypothesis is not true, but is easily done with the SAS macro (12).

For two factors, there are three relevant null hypotheses based on the normalized distributions, no main effect of A [$H_0^F(A)$], no main effect of B [$H_0^F(B)$], and no interaction of A and B [$H_0^F(AB)$]. These can be written as

$$\begin{aligned} H_0^F(A)&: \overline{F}_{1\bullet} = \overline{F}_{2\bullet} = \ldots = \overline{F}_{a\bullet} \\ H_0^F(B)&: \overline{F}_{\bullet 1} = \overline{F}_{\bullet 2} = \ldots = \overline{F}_{\bullet b} \\ H_0^F(AB)&: F_{ij} + \overline{F}_{\bullet\bullet} = \overline{F}_{i\bullet} + \overline{F}_{\bullet j} \end{aligned} \tag{11}$$

for all $i = 1, .., a$ and $j = 1, \ldots, b$. Note that $\overline{F}_{\bullet\bullet}$ is simply the mean (normalized) distribution across all levels of factors A and B. These hypotheses are all straightforward generalizations of the ones based on expected values ($\mu_{i\bullet}$, $\mu_{\bullet j}$, and $\mu_{\bullet\bullet}$) for a two-way ANOVA and presented in introductory textbooks on ANOVA and experimental design. General interpretation of the hypotheses is the same as with ANOVA, except that the response is the (normalized) distribution of the data and not just the expected value. For example, an interaction means that the effect of the *i*th level of A on the distribution depends on the level of B, or that the effect of the *j*th level of B depends on the level of A. It should be noted that

TABLE 1. Median, mean rank, and estimated relative treatment effects ($\hat{p}_i$) for the severity of foliar symptoms on potato caused by *Verticillium dahliae* isolates[a]

Isolate	Vegetative compatibility group	Median disease rating	Mean rank ($\overline{R}_{i\bullet}$)	Estimated relative effect ($\hat{p}_i$)[b]	Confidence interval (95%) for relative treatment effect	
					Lower limit	Upper limit
83	4A	3.5	42.8	0.88 (0.019)	0.82	0.90
111	4A	2.0	27.8	0.57 (0.051)	0.47	0.66
120	4A	2.0	21.6	0.44 (0.084)	0.29	0.60
201	4B	1.5	17.6	0.36 (0.076)	0.23	0.52
202	4B	1.0	11.6	0.23 (0.050)	0.16	0.36
203	4B	2.0	25.7	0.52 (0.066)	0.40	0.65

[a] Omer et al. (52).

[b] Standard errors (*se*) are given in brackets after the $\hat{p}_i$ estimates, which were determined based on output of the SAS macro. In general, if the SAS macro is not available, se($\hat{p}_i$) can be roughly approximated by $se(\overline{R}_{i\bullet})/N$, in which $se(\overline{R}_{i\bullet})$ is the standard error of the mean rank for the *i*th isolate (treatment) as determined in the MIXED procedure of SAS with the LSMEANS option (SAS Institute, Cary, NC).

there may be a significant interaction in terms of distributions even when there is no interaction in terms of means (18).

The main effects and interaction can be tested with both a WTS and an ATS, with the latter being preferred for typical sample sizes in plant pathology (say, 20 or fewer replications). Moreover, various other hypotheses can be tested using contrasts in the same way that was done for the one-way layout. Linear rank statistics also can be used to test hypotheses with specific patterned alternatives.

Two-way factorial example. Krause et al. (35) examined the suppressive effects of nine different potting mixes (potting mixes = 1, …, 9; actual descriptions are given in the cited article) on Rhizoctonia damping-off of radish seedlings. The mixes had been either fortified or not (natural) with a combination of the biocontrol organisms *Chryseobacterium gleum* 299 ($C_{299}R_2$) and *Trichoderma hamatum* 382 (T_{382}). Damping-off severity was determined 7 days after planting on a 1-to-5 ordinal scale: 1 = symptomless; 2 = small root or stem lesion; 3 = large root or stem lesion; 4 = postemergence damping-off; and 5 = preemergence damping off. Ratings were done on each of five pots (consisting of 32 plants each) per combination of potting mix and fortification, and there were eight separate batches (replications). The different pots served as subsamples, and the median rating over the five pots was calculated for each mix–fortification–replication combination. Brunner et al. (chapter 9 in literature citation 12) gives an alternative way of dealing with the subsamples. This is a two-factor crossed experimental design.

The median disease ratings, mean ranks, and estimated relative marginal effects are shown in Table 2. Ratings ranged from 1.2 up to 4.2, and estimated relative effects ranged from 0.14 up to 0.80, which clearly varied with potting mix and fortification with biocontrol agents. A quick scan through the $\hat{p}_{ij}$ values indicates that there were some large differences among the potting mixes in terms of their suppression of Rhizoctonia damping-off, but maybe not overall between fortified and unfortified mixes.

The test statistics are given in Table 3. Based on the ATS, there was a significant effect of potting mix on the suppression of damping-off ($P < 0.001$), but the main effect of mix fortification was only marginally significant ($P \approx 0.06$). However, there was a significant interaction of potting mix and fortification ($P = 0.023$), indicating that the difference between natural and fortified mixes depended on the specific potting mix. Based on the estimated relative effects and their standard errors, fortification significantly decreased disease in potting mixes 3 and 9 and not the others. Contrasts of the mean ranks confirmed that fortification was effective in these two mixes.

The same conclusions would be made in this example using the WTS or the ATS (Table 3). It should be noted that when there are only two levels of a factor, as with fortification here, the two test statistics are the same. The *P* values can be different, however, because a chi-square test is performed with the WTS and an *F* test is performed with the ATS. The main problem with using the WTS is that the test statistic is only asymptotically chi-square, and studies show that the asymptotic value is not approached rapidly with increasing sample size (11,13). Misleading (overly small) *P* values can be obtained at small-to-moderate sample sizes when the null hypothesis is actually true (11). The ATS is also based on asymptotic theory (11), but studies show the approximation is very accurate at most typical sample sizes (12). For any single data set, the *P* values for the WTS are smaller than the ATS. This can be seen with the test for the interaction of fortification and potting mix. Although there was no impact on conclusions here, it is very possible to falsely reject the null hypothesis using the WTS in borderline cases. For the remainder of the article, we refer only to results based on the ATS.

Repeated measures: concepts. It is common in plant pathology to assess disease severity repeatedly over time in the same experimental units (30,40,67). The data obtained are often referred to as repeated measures or longitudinal data. Experimental designs of this type are known as repeated measures designs. In one common approach to data analysis, assessment time is considered one of the factors in the experiment. However, a consequence of this method of data collection is that the data from the same experimental unit are correlated, but that data from different experimental units can still be assumed to be independent (45).

We consider only the two-factor situation here, but the approach is expandable to multiple factors. Factor A is a treatment-type factor (with *a* levels) that is being investigated, and factor B is the time factor (with $j = 1, \ldots, b$ levels). For simplicity, we refer to factor A as treatment. Almost all of the presentation on concepts

TABLE 2. Median, rank, and relative treatment effects for severity rating of Rhizoctonia damping-off of radish seedlings in relation to potting mixes and fortification with biological controls[a]

	Median disease rating		Mean rank ($\overline{R}_{ij\bullet}$)		Estimated relative treatment effect ($\hat{p}_{ij}$)[b]	
Potting mix	Natural	Fortified	Natural	Fortified	Natural[c]	Fortified
1	1.2	1.2	21.9	20.6	0.15 (0.030)	0.14 (0.027)
2	3.6	3.2	95.1	82.6	0.66 (0.049)	0.57 (0.049)
3	3.5	2.3	93.2	54.4	0.64 (0.076)*	0.37 (0.012)
4	1.1	1.3	23.4	29.9	0.16 (0.032)	0.20 (0.033)
5	4.0	4.2	107.9	115.9	0.75 (0.073)	0.80 (0.059)
6	3.7	3.9	105.6	103.1	0.73 (0.053)	0.71 (0.061)
7	1.2	1.2	25.3	26.8	0.17 (0.034)	0.18 (0.038)
8	3.8	3.8	100.8	108.6	0.70 (0.044)	0.75 (0.069)
9	3.9	3.0	109.4	80.4	0.76 (0.045)*	0.56 (0.057)

[a] Data from Table 2 of Krause et al. (35) and description of the potting mixes.
[b] Standard errors (*se*) are given in brackets after the $\hat{p}_i$ estimates, which were determined based on output of the SAS macro. In general, if the macro is not available, $se(\hat{p}_i)$ can be roughly approximated by $se(\overline{R}_{i\bullet})/N$, in which $se(\overline{R}_{i\bullet})$ is the standard error of the mean rank for the *i*th treatment as determined in the MIXED procedure of SAS with the LSMEANS option (SAS Institute, Cary, NC).
[c] Asterisk indicates that the relative effect is significantly different between natural potting mix and biocontrol-fortified mix ($P = 0.05$).

TABLE 3. Test statistics for the effects of potting mix and fortification with biological controls on Rhizoctonia damping-off of radish seedlings[a]

	Wald-type statistic				Analysis of variance-type statistic			
Effect	df_N	df_D	Chi-square	*P* value	df_N	df_D	*F*	*P* value
Fortification	1	126	3.54	0.060	1.00	89.1	3.54	0.063
Potting mix	8	126	529.28	<0.001	6.83	89.1	49.26	<0.001
Fortification × potting mix	8	126	18.84	0.016	6.83	89.1	2.49	0.023

[a] Data from Krause et al. (35). df_N = numerator degrees of freedom; and df_D = denominator degrees of freedom.

and methodology for the two-factor crossed design in the previous sections applies here and is not repeated. However, because there are not separate plots for each assessment time of each treatment, but just for treatments, the number of independent replications for each treatment is n_i (as with the one-way layout), and not n_{ij}. (Note that Brunner et al. [12] reverses the order of the subscripts for repeated measures, using X_{ikj} instead of X_{ijk}).

Because of the dependencies within an experimental unit (subject), if j and j' represent two different times, there is a nonzero correlation (or covariance) between measurements at these times within replicates (X_{ijk} and $X_{ij'k}$). The dependencies within experimental units can be directly handled using the relative marginal effects analysis. With linear models and normally distributed data, certain functional forms (e.g., autoregressive, antedependence) are commonly used for the correlation (or covariance) of X between times within experimental units (subjects) (59). However, with the nonparametric approach, one usually must assume a completely arbitrary covariance matrix for characterizing the variances of X at each time (at each level of A) as well as the covariances of X between all possible pairs of times (within each level of A). This is due, in part, to the nonlinear aspect of the rank transformation (2).

Due to the correlations, the statistical methodology of a repeated measures analysis is more complicated than a crossed factorial. Readers can refer to Wolfinger (66) and Lindsey (38) for more general information on this subject. Two features of the nonparametric analysis are worth pointing out here. First, the estimated relative marginal effects are not unbiased for small or moderate sample sizes (for example, less than 10 replications) (12). Fortunately, the bias can be estimated and used to make corrections. Second, the denominator degrees of freedom (df_D) for the ATS equals ∞ for the main effect of B (time) and the A–B interaction, rather than a finite number determined for crossed factorials.

Repeated measures: example. We revisit the data collected by Omer et al. (52) on PED. The data set examined here is the same as in the one-way layout example, but now we consider weekly disease assessments made over a 6-week period. So, the data file contains variables for isolate, time, rating, and subject (a unique identifier for each isolate–replication combination).

Median disease ratings and the estimated relative treatment effects ($\hat{p}_{ij}$) over time are shown in Figure 2. Disease severity increased over the 6-week assessment period for all isolates. Test statistics for the overall effects of isolate, time, and the isolate–time interaction are shown in Table 4. The significant interaction indicates that the isolates had different disease rating curves over time.

When there are more than two levels of the treatment factor (i.e., $a > 2$), one can test for so-called pairwise interactions. In particular, for every pair of treatments, i and i', the null hypothesis is tested that the difference in the (normalized) distributions is the same at all times. This can be written as

$$H_0^F(AB): F_{ij} - F_{i'j} = \overline{F}_{i\bullet} - \overline{F}_{i'\bullet} \quad \text{for } j = 1, \ldots, b \tag{12}$$

The equation for this hypothesis is just a special case of the third part of equation 11 (for interaction) when there are only two levels of factor A. A significant test result indicates that the pair of disease rating curves is different over time (i.e., that $p_{ij} - p_{i'j}$ is not constant at all j values). The ATSs for these global pairwise comparisons are shown in Table 5. Some profiles were similar (e.g., isolates 111 and 120), but many were significantly different in this example.

A question of interest at the start of the study could be whether some isolates are more aggressive than others. If so, one possible outcome would be disease progress curves which diverge over time. This is a more specific alternative hypothesis than the global one (tested previously) of any difference in progress curves over time. More specific hypotheses (patterned alternatives or interactions) concerning disease progress curves can be tested using the methodology of this paper. In this example, we tested the hypothesis that the difference in relative treatment effects between pairs of isolates ($p_{ij} - p_{i'j}$) increased linearly with time. The pattern is thus summarized as $\sum w_j(p_{ij} - p_{i'j})$, in which w_j = 1, 2, 3, 4, 5, and 6 are the weight coefficients for each week.

Test statistics are given in Table 5. P values for this patterned alternative test were lower than the P values for the global alternative for pairs of isolates with clearly diverging rating curves (e.g., isolate 111 versus 201), indicating the higher power of the patterned alternative test when the hypothesized alternative is true. On the other hand, for pairs of curves that were well separated (and significant based on the global test) but did not diverge over time (e.g., isolate 83 versus 201), the P value for the patterned alternative test was not significant (Table 5).

In separate analyses, we found that the marginal effects analysis could be use to analyze disease progress curves in field plots with as few as four replications (D. A. Shah and L. V. Madden, *unpublished data*).

Split plot. In a split-plot experimental design, one assigns the a levels of the first factor (e.g., factor A) to the larger experimental units (whole plots), which are then divided (or split up) into smaller units (subplots) to which one randomly assigns the b levels of the second factor (e.g., factor B) within each whole plot. In other words, the experimental unit for one of the factors (factor B) is a subunit of the experimental unit with respect to another factor (factor A) (59).

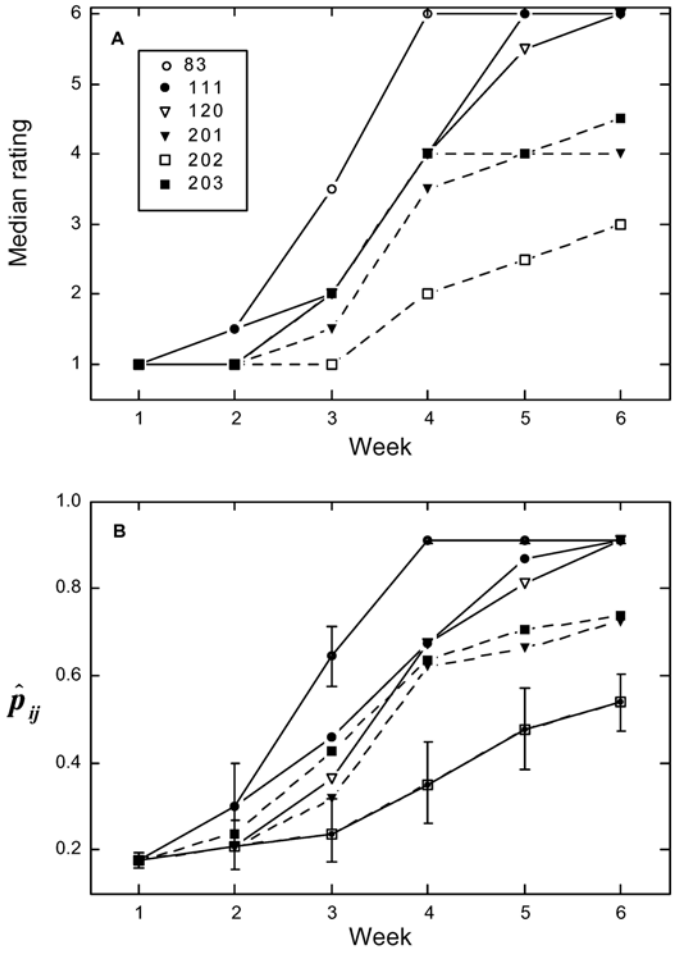

Fig. 2. A, Median disease ratings and **B,** estimated relative marginal effects ($\hat{p}_{ij}$) for the severity of potato early dying symptoms measured over a 6-week period (52). The 95% confidence intervals are shown for isolates 83 and 202 only (to avoid major overlap of intervals).

Although a split plot may appear different from the repeated measures design, the two have much in common and can be analyzed nonparametrically in much the same way. The splitting of each experimental unit imposes nonzero correlations (or covariances) between observations within each large experimental unit, that is, between X_{ijk} and $X_{ij'k}$. Although for linear models and normally distributed data, the correlations of X for a split plot can be assumed to be fixed for all pairs of j and j' (because the subplot factor levels are randomized within the whole-plot units), for the nonparametric analysis, the covariance structure could be more complicated (3). In particular, the variance at each time for each whole-plot factor level could be different. Thus, one can analyze data from split plots using the methods and dedicated software (macros) for repeated measures or with PROC MIXED of SAS. One simply substitutes the subplot factor for the time factor of a repeated measures analysis. This approach allows an unspecified variance–covariance matrix. Other variance–covariance structures could be considered using PROC MIXED of SAS. Using data published by Harveson and Rush (30), we have found that the nonparametric analysis was effective for testing for whole-plot, subplot, and interaction effects, and for comparing estimated relative marginal effects for the different factor levels (D. A. Shah and L.V. Madden, *unpublished data*).

DISCUSSION

Statisticians and other researchers often fall into two camps in terms of how quickly one should abandon parametric (typically, normal-distribution-based) methods in favor of nonparametric procedures. Researchers in the one camp will argue that the distributional assumptions of ANOVA and related methods can rarely, if ever, be achieved, and primarily rely on nonparametric statistical methods. Those in the other camp will point out that ANOVA is remarkably robust to moderate (or even greater) violations of the assumptions, and routinely use parametric methods. With the recent advances in the theory and application of mixed models to data analysis (45), even traditional problems such as nonconstant variance and complicated correlations of observations can be handled in a parametric framework. Moreover, generalized linear models (GLMs) and generalized linear mixed models (GLMMs) can be used to analyze, parametrically, data with common discrete distributions such as the binomial, Poisson, and negative binomial (41,44,45,55).

Parametric statistical methods can be used successfully for a wide range of data analysis problems. However, certain data measurement classes clearly pose serious problems for parametric analyses. Use of ANOVA for ordinal data is problematic, as discussed in the introduction. It has been straightforward for decades to properly test for the effects of treatments on biological responses using nonparametric methods, but these analyses generally were restricted primarily to simple one-way experimental designs or randomized complete block designs (with one fixed factor).

Consequently, researchers in many fields typically do one of four things when dealing with factorials. First, they ignore the problems of ordinal measurement scales and analyze the data using parametric methods. Such an approach has been common in the social sciences, where the data often consist of rating scores of behavior or conditions or the ordinal preference data of individuals (e.g., strongly agree versus strongly disagree; worst versus best) (28,47). As discussed by Snedecor and Cochran (63), for ANOVA to be appropriate for such ordinal (category) data, one must assume that the rating values represent equal gradations on an underlying (unobserved) scale. This assumption often is difficult or impossible to verify (59).

The second approach is to acknowledge the inherent properties of ordinal data (either implicitly or explicitly) but perform additional analyses to determine if the ordinal rating score behaves similarly to a continuous scale variable. (Even normality of a rating value can be assessed if there are enough observations, although the meaning may not be clear [47]). Then ANOVA is used on the rating data when such an approach seems reasonable. For instance, Lipps and Madden (39) showed that there was a very high correlation between a 0 to 10 rating for wheat powdery mildew (Table 2 in literature citation 39) and the directly estimated percent disease severity. This is a case where the assumption of equal gradations of an underlying disease severity scale

TABLE 4. Test statistics for the effects of *Verticillium dahliae* isolate and time after inoculation on the severity of potato early dying symptoms[a]

	Analysis of variance-type statistic			
Effect	df_N	df_D	F	P value
Isolate	4.22	34.13	35.28	<0.001
Time	3.13	∞	457.78	<0.001
Isolate × time	10.90	∞	7.78	<0.001

[a] Data from Omer et al. (52). df_N = numerator degrees of freedom; and df_D = denominator degrees of freedom.

TABLE 5. Tests of isolate and time interactions for pairs of isolates (i.e., pairwise comparisons) for potato early dying[a]

	Global alternatives[b]			Patterned alternative interaction[c]		
Isolate comparison	ATS	df_N	P value	L	df	P value
111 vs. 120	1.17	2.85	0.318	–0.74	9.63	0.761
111 vs. 201	3.40	2.79	0.019	3.55	13.52	0.002
111 vs. 202	11.08	2.59	<0.001	5.95	11.80	<0.001
111 vs. 203	2.89	2.17	0.051	3.09	13.87	0.004
111 vs. 83	7.53	2.07	<0.001	–0.59	12.01	0.718
120 vs. 201	3.51	2.82	0.017	5.78	10.71	<0.001
120 vs. 202	13.62	3.36	<0.001	7.26	8.08	<0.001
120 vs. 203	4.61	2.74	0.004	4.55	9.19	<0.001
120 vs. 83	9.90	2.77	<0.001	0.13	12.34	0.451
201 vs. 202	5.82	2.73	<0.001	3.67	10.53	0.002
201 vs. 203	0.93	2.66	0.416	–0.01	12.97	0.503
201 vs. 83	11.46	2.24	<0.001	–5.06	13.26	1.000
202 vs. 203	6.14	2.78	<0.001	–3.40	12.48	0.998
202 vs. 83	25.74	2.66	<0.001	–6.92	9.28	1.000
203 vs. 83	7.93	2.01	<0.001	–4.14	11.33	0.999

[a] Data from test 1 of Omer et al. (52).

[b] ATS is the analysis of variance-type statistic (ATS) for the interaction of the two listed isolates and time, which has an approximate F distribution under the null hypothesis of no interaction. df_N = numerator degrees of freedom. The denominator degrees of freedom (df_D) is ∞, and not listed. The alternative hypothesis is any type of interaction (i.e., global alternative).

[c] L is the test statistic for the patterned interaction of the two listed isolates and time, which has an approximate t distribution under the null hypothesis of no interaction; df is the associated estimated degrees of freedom (determined using the formula on page 143 of Brunner et al. [12]).

may be warranted. In some instances, considerable effort has been invested in relating rating data to continuous scale data (51,62). Moreover, researchers sometimes transform rating data to develop a disease score (commonly called a disease severity index) that is analogous to a continuous scale variable (10,19,30,35) with a normal distribution. The transformed ratings are then analyzed with ANOVA. These efforts may avoid or circumvent some problems associated with the parametric analysis of ordinal data. However, because parametric methods are not scale invariant, interpretation of the results may be ambiguous, especially if the effects of the transformation (rating scale compression, elongation, or introduction of nonlinearity) are unknown.

A third possible approach used by researchers is to ignore the factorial structure of the experimental design and analyze the data as if they originated from a one-way layout (25). For instance, if the design was a crossed three-way factorial, with two, three, and four levels of factors A, B, and C, then one creates a single pseudo-treatment factor (τ) with $2 \times 3 \times 4 = 24$ levels and determines the effect of treatment on the response using the Kruskal-Wallis (37) procedure or other appropriate approach. Such an approach is equivalent to assuming a three-way interaction, which may or may not be justified, because there are no explicit tests for main effects or any of the interactions. Moreover, this approach cannot be used for split plots or repeated measures because the correlations of observations are not accounted for.

The fourth approach consists of ranking the data and then simply using ANOVA on the rank-transformed data. This rank transform method (23,24) is common in some fields (58) and has been advocated by some statisticians (24) as a nonparametric method. The statistical problems with this approach were summarized in the introduction and are considered in detail elsewhere (2,7,8,16,64,65). The statistical evidence is now strong that the rank transform method should not be used as a general method of data analysis because of incorrect results for tests of many different hypotheses of interest regarding the means with factorial designs.

With the recent advances in the theory and application of relative marginal effects and hypothesis testing of distribution functions for factorial layouts (4–6,12,13,17,18), researchers now have some useful and statistically sound nonparametric alternatives to the less-than-desirable methods described previously for analyzing ordinal data. Although parametric approaches certainly offer the most flexibility in analyzing data from the full range of experimental designs, as demonstrated here, factorials, split plots, and repeated measures can all be analyzed appropriately in a nonparametric framework with disease rating and other ordinal-scale data (12,17). In fact, the approach can also be taken for continuous data in which the distributional assumptions of parametric analysis are not justified. The new nonparametric analyses can be done using commercially available software from SAS and using free macros for the SAS and R statistical systems. Main effects, interactions, and other contrasts can be tested with WTS or ATS of the ranks, the latter being preferred for the typical number of replications in most studies. Contrasts can be constructed for testing meaningful hypotheses, with either so-called patterned alternatives or global alternatives (Table 5). Main effects and their interactions can be quantified by estimated relative marginal effects, $\hat{p}_{ij}$, and the (estimated) standard errors of $\hat{p}_{ij}$, which are directly related to the ranks of the observations. As stated by Brunner et al. (12) in a repeated-measures context, "not only are these relative effects independent of the specific choice of the grading [rating] scale, … but they also allow smaller tendencies in the time curves to be depicted."

A desirable property of ANOVA, if assumptions are (reasonably) met, is that the procedure can be used for very small number of replications (e.g., three), although the power of the tests and the precision of the parameter estimates (i.e., standard error of expected values for each treatment level) increases dramatically with increasing number of replications. Asymptotic theory is used with the nonparametric marginal effects approach to derive the Wald- and ANOVA-type test statistics (WTS and ATS) and determine the distribution of the estimated relative effects when the null hypothesis is not true (18). However, the approach using the ATS works well with moderate or even small sample sizes (14,26). Moreover, Brunner et al. (11) have shown that the power of the marginal effects analysis is high under many circumstances and is similar to ANOVA for normally distributed data. Brunner and colleagues do recommend more than the three replications that are often used in agricultural experiments (13); however, a large number of replications will not always be practical. In preliminary numerical studies, we found that the approximation of the ATS with an F statistic is reasonable even with as few as four replications (L. V. Madden, *unpublished data*). If one has the choice between a large number of replications of small size or a small number of replications of large size (i.e., with many subsamples), then the former is preferred for the nonparametric analysis. The exception would be if plot size directly determines the results (20,40) because of, for example, disease spread within and between plots. In this case, finding a way to use a continuous scale for assessing disease would be preferable.

Although this article argues heavily in favor of using the marginal effects analysis (12,17), there are some valuable alternatives when the original data are ordinal in scale. Under some circumstances, it is useful to use randomization and resampling methods to test hypotheses about the effects of factors (43), although we find these methods to be of most use, both in terms of interpretation of results and in carrying out the analysis, when there are only a (relatively) small number of factors and levels to the factors (L. V. Madden and D. A. Shah, *unpublished data*). It should be pointed out that some randomization and resampling methods still treat data as continuous rather than ordinal. If the ordered values are recorded on an integer scale, one can determine the proportion of observations greater than (or less than) a specific integer. For instance, with a 1-to-4 scale, one could define the variable y as 1 if the rating is >2, and 0 otherwise. The fraction of individuals across the replications with $y = 1$ is then an indication of the severity of the disease for the population. If there is one observation per combination of factor levels and replications, the new variable is binary (either 0 or 1), and GLMs (e.g., logistic model analysis) can be used to test for main effect and interactions (22). If there are subsamples for each combination of factor level and replication, then the proportion y/n (with n = number of subsamples) can be analyzed with GLMs, GLMMs, or with ANOVA after suitable transformation of the proportions (32,33,41,55,68). In fact, Krause et al. (35) took this approach, in part, in analyzing their rating data. The approach can be extended to encompass the entire range of rating categories and analyzed with ordered logistic models (1,59). However, it may be difficult to fit ordered logistic models to data without a large number of replications if there are many different rating categories, many levels of the experimental factors, or many different factors. Nevertheless, considerable progress is being made in this area of data analysis (1) that will be of direct value to plant pathologists.

Continuous scale data are very useful for developing functional relationships between disease intensity and time, space, and other variables, based on our understanding of the biological mechanisms and (current and past) empirical evidence (20). The marginal effects nonparametric analysis discussed here does allow for some quantification of trends of ordinal data over time or space through an analysis and comparison of the estimated relative marginal effects. With proper choice of contrasts, with or without the use of patterned alternative hypotheses, relatively sophisticated analysis of disease increase in time and spread in space can be accomplished. Nevertheless, the approach does not fully permit the determination of a functional relationship between disease intensity and the variables of interest. This is a limitation of ordinal

data, not of the statistical methods discussed. Thus, continuous-scale measurements of disease are advantageous for many quantitative epidemiological studies. Conversely, continuous-scale measurements of disease are only useful if severity can be determined accurately and reliably (50,53). Ordinal measurements are beneficial for diseases that are difficult to measure quantitatively (such as those caused by soilborne pathogens and by viruses). For these situations, the approach discussed in this paper will be advantageous.

ACKNOWLEDGMENTS

Salaries and research support were provided by state and federal funds appropriated to the Ohio Agricultural Research and Development Center, The Ohio State University. We thank B. Harveson, M. Omer, R. Rowe, and P. Lipps for making their data available to us and E. Brunner for very helpful advice.

LITERATURE CITED

1. Agresti, A., and Natarajan, R. 2001. Modeling clustered ordered categorical data: A survey. Int. Stat. Rev. 69:345-371.
2. Akritas, M. G. 1990. The rank transform method in some two-factor designs. J. Am. Stat. Assoc. 85:73-78.
3. Akritas, M. G. 1991. Limitations of the rank transform procedure: A study of repeated measures designs, Part I. J. Am. Stat. Assoc. 86:457-460.
4. Akritas, M. G., and Arnold, S. F. 1994. Fully nonparametric hypotheses for factorial designs I: Multivariate repeated measures designs. J. Am. Stat. Assoc. 89:336-343.
5. Akritas, M. G., Arnold, S. F., and Brunner, E. 1997. Nonparametric hypotheses and rank statistics for unbalanced factorial designs. J. Am. Stat. Assoc. 92:258-265.
6. Akritas, M. G., and Brunner, E. 1997. A unified approach to rank tests for mixed models. J. Stat. Plan. Infer. 61:249-277.
7. Akritas, M. G., and Osgood, D. W. 2002. Guest editors' introduction to the special issue on nonparametric models. Sociol. Methods Res. 30:303-308.
8. Blair, R. C., Sawilowsky, S. S., and Higgins, J. J. 1987. Limitations of the rank transform statistic in test for interactions. Commun. Stat. Part B Simul. Comput. 16:1133-1145.
9. Box, G. E. P. 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. Ann. Math. Stat. 25:290-302.
10. Bradley, C. A., Hartman, G. L., Wax, L. M., and Pedersen, W. L. 2002. Influence of herbicides on Rhizoctonia root and hypocotyl rot of soybean. Crop Prot. 21:679-687.
11. Brunner, E., Dette, H., and Munk, A. 1997. Box-type approximations in nonparametric factorial designs. J. Am. Stat. Assoc. 92:1494-1502.
12. Brunner, E., Domhof, S., and Langer, F. 2002. Nonparametric Analysis of Longitudinal Data in Factorial Experiments. John Wiley & Sons, New York.
13. Brunner, E., Domhof, S., and Puri, M. L. 2002. Weighted rank statistics in factorial designs with fixed effects. Stat. Neerl. 56:179-194.
14. Brunner, E., and Munzel, U. 2000. The nonparametric Behrens-Fisher problem: Asymptotic theory and a small-sample approximation. Biom. J. 42:17-25.
15. Brunner, E., and Munzel, U. 2002. Nichtparametrische Datenanalysen: Unverbundene Stichproben. Springer-Verlag, Berlin.
16. Brunner, E., and Neumann, N. 1986. Rank tests in 2×2 designs. Stat. Neerl. 40:251-271.
17. Brunner, E., and Puri, M. L. 2001. Nonparametric methods in factorial designs. Stat. Pap. 42:1-52.
18. Brunner, E., and Puri, M. L. 2002. A class of rank-score tests in factorial designs. J. Stat. Plan. Infer. 103:331-360.
19. Bruton, B. D., Garcia-Jimenez, J., Armengol, J., and Popham, T. W. 2000. Assessment of virulence of *Acremonium cucurbitacearum* and *Monosporascus cannonballus* on *Cucumis melo*. Plant Dis. 84:907-913.
20. Campbell, C. L., and Madden, L. V. 1990. Introduction to Plant Disease Epidemiology. John Wiley & Sons, New York.
21. Cintas, N. A., and Webster, R. K. 2001. Effects of rice straw management on *Sclerotium oryzae* inoculum, stem rot severity, and yield of rice in California. Plant Dis. 85:1140-1144.
22. Collett, D. 2002. Modelling Binary Data. 2nd ed. Chapman & Hall, London.
23. Conover, W. J. 1998. Practical Nonparametric Statistics. 3rd ed. John Wiley & Sons, New York.
24. Conover, W. J., and Iman, R. L. 1981. Rank transformations as a bridge between parametric and nonparametric statistics (with discussion). Am. Stat. 35:124-133.
25. Craft, C. M., and Nelson, E. B. 1996. Microbial properties of composts that suppress damping-off and root rot of creeping bentgrass caused by *Pythium graminicola*. Appl. Environ. Microbiol. 62:1550-1557.
26. Delaney, H. D., and Vargha, A. 2002. Comparing several robust tests of stochastic equality with ordinally scaled variables and small to moderate sized samples. Psychol. Methods 7:485-503.
27. Edwards, S. J., Collin, H. A., and Isaac, S. 1997. The response of different celery genotypes to infection by *Septoria apiicola*. Plant Pathol. 46:264-270.
28. Gould, J. E. 2002. Concise Handbook of Experimental Methods for the Behavioral and Biological Sciences. CRC Press, Boca Raton, FL.
29. Harikrishnan, R., and Yang, X. B. 2002. Effects of herbicides on root rot and damping-off caused by *Rhizoctonia solani* in glyphosate-tolerant soybean. Plant Dis. 86:1369-1373.
30. Harveson, R. M., and Rush, C. M. 2002. The influence of irrigation frequency and cultivar blends on the severity of multiple root diseases in sugar beets. Plant Dis. 86:901-908.
31. Hora, S., and Conover, W. 1984. The *F*-statistic in the two-way layout with rank-score transformed data. J. Am. Stat. Assoc. 79:668-673.
32. Hughes, G., Munkvold, G. P., and Samita, S. 1998. Application of the logistic-normal-binomial distribution to the analysis of Eutypa dieback disease incidence. Int. J. Pest Manage. 44:35-42.
33. Hughes, G., and Samita, S. 1998. Analysis of patterns of pineapple mealybug wilt disease in Sri Lanka. Plant Dis. 82:885-890.
34. Khan, T. N., and Boyd, W. J. R. 1969. Physiologic specialisation in *Drechslera teres*. Aust. J. Biol. Sci. 22:1229-1235.
35. Krause, M. S., Madden, L. V., and Hoitink, H. A. J. 2001. Effect of potting mix microbial carrying capacity on biological control of Rhizoctonia damping-off of radish and Rhizoctonia crown and root rot of Poinsettia. Phytopathology 91:1116-1123.
36. Kruskal, W. H. 1952. A nonparametric test for the several sample problem. Ann. Math. Stat. 23:525-540.
37. Kruskal, W. H., and Wallis, W. A. 1952. Use of ranks in one-criterion variance analysis. J. Am. Stat. Assoc. 47:583-621.
38. Lindsey, J. K. 1999. Models for Repeated Measurements. 2nd ed. Oxford University Press, Oxford.
39. Lipps, P. E., and Madden, L. V. 1989. Assessment of methods of determining powdery mildew severity in relation to grain yield of winter wheat cultivars in Ohio. Phytopathology 79:462-470.
40. Lipps, P. E., and Madden, L. V. 1992. Effects of plot size and border width on assessment of powdery mildew of winter wheat. Plant Dis. 76:299-303.
41. Madden, L. V., Turechek, W. W., and Nita, M. 2002. Evaluation of generalized linear mixed models for analyzing disease incidence data obtained in designed experiments. Plant Dis. 86:316-325.
42. Magee, J. B., Smith, B. J., and Rimando, A. 2002. Resveratrol content of muscadine berries is affected by disease control spray program. Hortscience 37:358-361.
43. Manly, B. F. J. 1997. Randomization, Bootstrap and Monte Carlo Methods in Biology. Chapman & Hall, London.
44. McCullagh, P., and Nelder, J. A. 1989. Generalized Linear Models. CRC Press, Boca Raton, FL.
45. McCulloch, C. E., and Searle, S. R. 2001. Generalized, Linear, and Mixed Models. John Wiley & Sons, New York.
46. McSpadden-Gardener, B. B., and Lilley, A. K. 1997. Application of common statistical tools. Pages 501-523 in: Modern Soil Microbiology. J. D. van Elsas, J. T. Trevors, and E. M. H. Wellington, eds. Marcel Dekker, New York.
47. Munzel, U., and Bandelow, B. 1998. The use of parametric vs. nonparametric tests in the statistical evaluation of rating scales. Pharmacopsychiatry 31:222-224.
48. Murphy, J., Zehnder, G., Schuster, D., Sikora, E., Polstron, J., and Kloepper, J. 2000. Plant growth-promoting rhizobacterial mediated protection in tomato against *Tomato mottle virus*. Plant Dis. 84:779-784.
49. Nutter, F. W., Jr. 1997. Quantifying the temporal dynamics of plant virus epidemics: A review. Crop Prot. 16:603-618.
50. Nutter, F. W., Jr., Gleason, M. L., Jenco, J. H., and Christians, N. C. 1993. Assessing the accuracy, intra-rater repeatability, and interrater reliability of disease assessment systems. Phytopathology 83:806-812.
51. O'Brien, R. D., and van Bruggen, A. H. C. 1992. Accuracy, precision, and correlation to yield loss of disease severity scales for corky root of lettuce. Phytopathology 82:91-96.
52. Omer, M. A., Johnson, D. A., and Rowe, R. C. 2000. Recovery of *Verticillium dahliae* from North American certified seed potatoes and characterization of strains by vegetative compatibility and aggressiveness. Am. J. Potato Res. 77:325-331.

53. Parker, S. R., Whelan, M. J., and Royle, D. J. 1995. Reliable measurement of disease severity. Aspects Appl. Biol. 43:205-214.
54. Piccinni, G., and Rush, C. M. 2000. Determination of optimum irrigation regime and water use efficiency of sugar beet grown in pathogen-infested soil. Plant Dis. 84:1067-1072.
55. Piepho, H.-P. 1999. Analysing disease incidence data from designed experiments by generalized linear mixed models. Plant Pathol. 48:668-674.
56. Reid, L. M., Nicol, R. W., Ouellet, T., Savard, M., Miller, J. D., Young, J. C., Stewart, D. W., and Schaafsma, A. W. 1999. Interaction of *Fusarium graminearum* and *F. moniliforme* in maize ears: Disease progress, fungal biomass, and mycotoxin accumulation. Phytopathology 89:1028-1037.
57. Saari, E. E., and Prescott, J. M. 1975. A scale for appraising the foliar intensity of wheat diseases. Plant Dis. Rep. 59:377-380.
58. Sawilowsky, S. S. 2000. Review of the rank transform in designed experiments. Percept. Motor Skill 90:489-497.
59. Schabenberger, O., and Pierce, F. J. 2002. Contemporary Statistical Models for the Plant and Soil Sciences. CRC Press, Boca Raton, FL.
60. Scheiner, S. M., and Gurevitch, J. 2001. Design and Analysis of Ecological Experiments. 2nd ed. Oxford University Press, New York.
61. Scott, P. R., and Hollins, T. W. 1974. Effects of eyespot on the yield of winter wheat. Ann. Appl. Biol. 78:269-279.
62. Slopek, S. W. 1989. An improved method of estimating percent leaf area diseased using a 1 to 5 disease assessment scale. Can. J. Plant Pathol. 11:381-387.
63. Snedecor, G. W., and Cochran, W. G. 1989. Statistical Methods. 8th ed. Iowa State University Press, Ames.
64. Thompson, G. L. 1991. A unified approach to rank tests for multivariate and repeated measure designs. J. Am. Stat. Assoc. 86:410-419.
65. Thompson, G. L., and Ammann, L. P. 1990. Efficiencies of interblock rank statistics for repeated measures designs. J. Am. Stat. Assoc. 85:519-528.
66. Wolfinger, R. D. 1996. Heterogeneous variance-covariance structures for repeated measures. J. Agric. Biol. Environ. Stat. 1:205-230.
67. Xiao, C. L., and Subbarao, K. V. 2000. Effects of irrigation and *Verticillium dahliae* on cauliflower root and shoot growth dynamics. Phytopathology 90:995-1004.
68. Zarnoch, S. J., Anderson, R. L., and Sheffield, R. M. 1995. Using the beta-binomial distribution to characterize forest health. Can. J. For. Res. 25:462-469.