Symposium

New Applications of Statistical Tools in Plant Pathology

Nonparametric Tests in Plant Disease Epidemiology: Characterizing Disease Associations

W. W. Turechek

Department of Plant Pathology, New York State Agricultural Experiment Station, Cornell University, Geneva 14456.
Accepted for publication 12 May 2004.

ABSTRACT

Turechek, W. W. 2004. Nonparametric tests in plant disease epidemiology: Characterizing disease associations. Phytopathology 94:1018-1021.

Nonparametric tests are suited to many statistical applications, including experimental design, regression, and time series analysis, for example. Often these tests are thought of as alternatives to their parametric counterparts when certain assumptions about the underlying population are questionable. Although suited for this scenario, there are a number of nonparametric tests that fill unique niches in the analysis of data, for example, characterizing interspecific associations. Quantifying the degree of association between two or more pathogens or diseases at a defined spatial scale is essential to gain a thorough understanding of disease dynamics, generate testable hypothesis behind the mechanisms that cause association, and is often necessary in modeling applications. In this paper, nonparametric approaches to characterizing interspecific associations will be covered. Specifically, I will address the use of rank correlation coefficients and the development of a randomization procedure for testing the Jaccard index of association against a null model.

According to Hollander and Wolfe (5) "A nonparametric procedure is a statistical procedure that has certain desirable properties that hold under relatively mild assumptions regarding the underlying populations from which the data are obtained." Nonparametric tests have several desirable properties including (i) no assumption that the underlying population follows a normal distribution; (ii) they are typically easier to apply than their parametric counterparts; (iii) they are often easy to understand; (iv) they are usually only slightly less efficient than their normal counterparts when the underlying distribution is normal; and (v) they are relatively insensitive to outliers.

Nonparametric statistical tests are often thought of as "alternatives" to their parametric counterparts when data do not conform readily to the assumptions of standard parametric tests. Although this is true to some degree, nonparametric procedures should not be thought of merely as alternatives to parametrical tests. There are nonparametric tests well suited to a number of applications including experimental design, regression, and time series analysis. Indeed, there are a number of procedures that are uniquely nonparametric (e.g., randomizations and resampling statistics) that offer the analyst the tools needed to appropriately analyze their data where no parametric approach exists.

In this paper, nonparametric tests of independence will be covered. Tests of independence have many applications in the biological sciences, including their use for quantifying interspecific (or species) associations. In plant pathology, these tests could be used to measure the degree to which two or more pathogens (10), or two or more diseases (13), are associated. Associations can be characterized based on two underlying properties: covariation and occurrence (7). Covariation is a measure of how one disease's intensity (incidence or severity) increases or decreases in response to a change in intensity of another disease. Occurrence measures the degree to which two (or more) diseases occupy the same habitat (e.g., leaf, plant, field, region, etc.). Each property and the tests to measure them will be discussed in turn.

Corresponding author: W. W. Turechek; E-mail address: wwt3@cornell.edu

Publication no. P-2004-0719-05O

Quantifying covariation. The property of covariation is characterized typically with correlation coefficients. Correlation coefficients are used to measure the strength of a relationship between two variables when neither variable can be assumed to be the "explanatory" variable. Pearson's product moment correlation is perhaps the most common correlation coefficient and is applied appropriately when it can be assumed that a linear relationship exists between the two variables and that the variables are distributed according to a bivariate normal distribution. When these assumptions are not met or cannot be assumed, Pearson's product moment correlation should not be used and a nonparametric alternative should be sought.

The two most commonly used nonparametric or so-called "rank correlation coefficients" are Kendall's tau (τ) and Spearman's rho (ρ). Kendall's τ is derived from the closely related Kendall's statistic (K). A symmetric confidence interval about τ is obtained easily from K, but the opportunity is taken here to also develop an alternative confidence interval through an application of Efron's bootstrap. Spearman's ρ is introduced next. This statistic is calculated exactly as Pearson's product moment correlation except that the calculation is applied to rank transformed data rather than to the original. Spearman's ρ is then contrasted to Kendall's τ in the concluding remarks.

Kendall's test of concordance. Kendall's test of concordance requires the data to consist of n bivariate observations $(X_1,Y_1),\ldots,(X_n,Y_n)$, one observation on each of n subjects and assumes that the n (X,Y) pairs are mutually independent and identically distributed according to some continuous bivariate distribution. The test can be expanded to the multivariate case, but this will not be covered here.

Kendall's test of concordance tests the null hypothesis that the X and Y variables are independent. The test statistic, K, is calculated as $K = K' - K''$, where K' = the number of concordant pairs and

K'' = the number of discordant pairs. An (X,Y) pair is said to be concordant if $(X_i - X_j)(Y_i - Y_j) > 0$, and discordant if $(X_i - X_j)(Y_i - Y_j) < 0$. This approach to calculating K is applicable when no ties exist among the X's and no ties exist among the Y's. Variations for calculating τ that account explicitly for ties exist (e.g., τ_B and τ_C); these will not be covered here. A simpler method for calculating K, however, is to simply count the number of concordant pairs (K'). This is easily done by ordering the pairs from lowest to highest by their X values and then counting the number of pairs for which the corresponding Y's are in increasing order. The test statistic is then calculated: $K = 2K' - n(n-1)/2$.

Either a one or two-sided test of τ (or equivalently K) is possible by comparing the calculated value of K to the appropriate critical value, k_α, where k_α is chosen to make the type I error α. Tabulated (critical) values of k_α can be obtained from appendices in several nonparametric statistics textbooks for select sample sizes (n).

The test statistic K can be used to derive a correlation coefficient known as Kendall's sample rank correlation coefficient or Kendall's τ. Like most correlation coefficients, Kendall's τ assumes a value between 1 and –1. The correlation coefficient can be calculated using $\tau = 2K/[n(n-1)]$. The upper and lower bounds of a symmetric confidence interval around τ with confidence coefficient $(1-\alpha)$ are calculated using

$$\tau_L = \hat{\tau} - z_{\alpha/2}\hat{\sigma}, \quad \tau_U = \hat{\tau} + z_{\alpha/2}\hat{\sigma} \tag{1}$$

where $z_{\alpha/2}$ is the value from a standard normal distribution such that $\Pr(Z \geq z) = \alpha/2$ and

$$\hat{\sigma}^2 = \frac{2}{n(n-1)}\left[\frac{2(n-2)}{n(n-1)^2}\sum_{i=1}^{n}(C_i - \overline{C})^2 + 1 - \hat{\tau}^2\right] \tag{2}$$

$$C_i = \sum_{\substack{t=1 \\ t\neq i}}^{n} Q[(X_i,Y_i),(X_t,Y_t)], \text{ for } i = 1,\ldots,n, \tag{3}$$

$$Q[(a,b),(c,d)] = \begin{cases} 1, & \text{if } (d-b)(c-a) > 0 \\ -1, & \text{if } (d-b)(c-a) < 0 \end{cases} \tag{4}$$

and

$$\overline{C} = \frac{1}{n}\sum_{i=1}^{n} C_i = 2K/n \tag{5}$$

Because the confidence interval is based on a normal approximation, large sample sizes ($n > 30$) are generally required to assure the assumptions of the central limit theorem are met.

Efron's bootstrap. The confidence interval for Kendall's τ developed above was based on obtaining a mathematical expression for the standard deviation for τ. For some statistics, however, it is difficult or impossible to obtain an expression for the standard deviation. Efron's bootstrap is a general method for obtaining estimated standard deviations of statistics (estimators) and confidence intervals for parameters without requiring a tractable expression for the standard deviation (2).

Following Hollander and Wolfe (5), Efron's bootstrap can be used to calculate a confidence interval as follows: (i) Denote the bivariate sample as $Z_1 = (X_1,Y_1)$, $Z_2 = (X_2,Y_2),\ldots,$ $Z_n = (X_n,Y_n)$. (ii) Make n random draws with replacement from the bivariate sample. This is equivalent to taking an independent random sample from the n pairs. A possible bootstrap sample of a data set with $n = 11$ may contain 1 copy of Z_1, 3 copies of Z_2, 0 copies of Z_3, 0 copies of Z_4, 1 copy of Z_5, 3 copies of Z_6, 1 copy of Z_7, and 1 copy each of Z_8 and Z_9, with each Z_i being drawn with probability $1/n$. Repeat step two B times. B should be a minimum of 100, but values of B closer to 1,000 (or even 10,000) are preferred. (iii) For each of the B draws compute τ. The B values can be denoted as $\hat{\tau}^1,\hat{\tau}^2,\ldots,\hat{\tau}^B$. These are called the bootstrap replications. (iv) Order the bootstrap values from smallest to largest: $\hat{\tau}^{(1)},\hat{\tau}^{(2)},\ldots,\hat{\tau}^{(B)}$. A distribution-free confidence interval with approximate confidence coefficient $100(1-\alpha)\%$ is

$$\tau_L = \hat{\tau}^{(k)}, \quad \tau_U = \hat{\tau}^{(B+1-k)} \tag{6}$$

where $k = B(\alpha/2)$. For example, if $\alpha = 0.1$ and $B = 1{,}000$, $k = 1{,}000(0.05) = 50$, the lower and upper bounds of the interval are the $\hat{\tau}^{(50)}$ and $\hat{\tau}^{(951)}$ observations of the ordered bootstrap replications, respectively. The bootstrap estimated standard error is

$$\hat{\sigma}_B = \left\{\frac{\sum_{i=1}^{B}(\hat{\tau}^i - \overline{\tau})^2}{B-1}\right\}^{1/2} \tag{7}$$

where

$$\overline{\tau} = \frac{\sum_{i=1}^{B}\hat{\tau}_i}{B} \tag{8}$$

Spearman's rank correlation coefficient. Spearman's rank correlation coefficient (ρ) is another statistic used routinely for quantifying the property of covariation between two species. Like Kendall's τ, values of ρ assume a number between 1 and –1. For Spearman's rank correlation, data should be composed of n bivariate observations $(X_1,Y_1),\ldots,(X_n,Y_n)$, one observation on each of n subjects. Use of the test assumes that the (X,Y) pairs are mutually independent and identically distributed according to some continuous bivariate distribution. To compute Spearman's rank correlation coefficient, order the n X observations from least to greatest and let R_i denote the rank of X_i. Similarly, order the n Y observations from least to greatest and let S_i denote the rank of Y_i. Then,

$$\rho = \frac{12\sum_{i=1}^{n}\left(\left(R_i - \frac{n+1}{2}\right)\left(S_i - \frac{n+1}{2}\right)\right)}{n(n^2-1)} = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n(n^2-1)} \tag{9}$$

where, $D_i = R_i - S_i$.

Like τ, either a one or two-sided test of ρ is possible by comparing the calculated value of ρ to the appropriate critical value, ρ_α, where ρ_α is chosen to make the type I error α. Critical values of ρ_α can be obtained in appendices of several nonparametric statistical textbooks.

Quantifying occurrence. There are a number of simple ecological indices useful for characterizing the property of occurrence (6). One of the most widely used indices is the Jaccard index of association, J. Values of the Jaccard index range from 0 to unity, where values close to 1 are indicative of a high degree of association and values close to 0 are indicative of a low degree of association or, essentially, dissociation. The index is written as $J = a/(a + b + c)$, where a represents the number of sampling units where both species (i.e., pathogens or diseases) occur, b represents the number of sampling units where only disease 1 is present, and c represents the number of sampling units where only disease 2 is present. The index simply represents the probability of both diseases occurring together in the population of sampling units where either disease occurs.

Recently, Turechek and Madden (13) developed a nonparametric statistical procedure to test observed values of J against a "null value" of J, or in other words, against the expected value of J under the null hypothesis of no association. A randomization procedure was developed to derive a sampling distribution for J and the expected value of the index ($\overline{J}_{ran}$) was calculated from this empirical sampling distribution. The jackknife procedure was used to estimate the standard error of J under the null hypothesis of no association. Each procedure will be covered in turn.

Calculating the expected value of J. Let $X = [x_1,\ldots,x_i,\ldots,x_n]'$ and $Y = [y_1,\ldots,y_i,\ldots,y_n]'$ represent vectors of length n (i.e., the number of sampling units), in which x_i and y_i are binary variables representing the presence (1) or absence (0) of the component diseases in sampling unit i and the prime symbol denotes the ma-

trix transpose operator. Thus, X and Y are equal length vectors of 0's and 1's representing the presence and absence of the two diseases of concern in a single data set. The n values of X and the n values of Y are separately rearranged at random, and the Jaccard index is calculated for the randomized data (J_j). This is repeated k times for a single data set. The mean or expected value of the Jaccard index is calculated using

$$\bar{J}_{ran} = \sum_{j=1}^{k} J_j / k \tag{10}$$

That is, the value $\bar{J}_{ran}$ represents the value of the index one would expect if the two component diseases were distributed independently given their observed incidences.

The jackknife. A nonparametric estimate of the standard error of J can be obtained using the jackknife procedure (1). Following Turechek and Madden (13), the jackknifed standard error can be calculated as follows: (i) calculate the standard Jaccard coefficient (J) as described above; (ii) remove the first observation and recalculate the coefficient based on the remaining data points (J_{-1}); (iii) repeat step two for each observation in turn in order to calculate n different J_{-i} values; and (iv) calculate of the ith pseudovalue (υ_i) for each observation as $\upsilon_i = J + (n - 1)(J - J_{-i})$. The jackknifed standard error of the Jaccard coefficient is then

$$s_j = \sqrt{\frac{\sum(\upsilon_i - \overline{\upsilon})^2}{n(n-1)}} \tag{11}$$

A test for association. To test whether the observed Jaccard index is significantly different from the value calculated under the assumption of independence, a normal distribution is assumed for the index estimated by J (the observed Jaccard value) with standard error estimated by the jackknifed value, s_j (8). A standard normal statistic can then be calculated using

$$Z = \frac{J - \bar{J}_{ran}}{s_j} \tag{12}$$

Treating Z as a two-sided test, values of $Z > 1.96$ indicate significant positive association and values of $Z < -1.96$ indicate significant negative association or dissociation at $P = 0.05$.

Conclusion. The nonparametric approach taken here for characterizing disease or interspecific associations draws upon a rather large set of tools used by ecologists of which only a few were demonstrated. Rank correlation statistics are used routinely to characterize the property of covariation. Although useful for characterizing this property, rank correlation statistics have other applications as well. For example, Turechek and Stevenson (14) used Kendall's sample rank correlation coefficient to measure the degree of association between components of partial resistance to pecan scab caused by *Cladosporium caryigenum*.

The most common question surrounding the use of these two correlation coefficients is "which of these should I use?" Spearman's ρ and Kendall's τ are, in general, measuring the same property but imply different interpretations: ρ can be thought of as the regular Pearson's product moment correlation coefficient applied to ranks. That is, ρ represents the proportion of variability that can be attributed to association between diseases. Kendall's τ represents a probability; it is the difference between the probabilities that two variables are in the same order versus that they are in different order (4,10).

In general, the two statistics will seldom lead to different conclusions when applied to the same set of data. Spearman's ρ is more sensitive to outliers than Kendall's τ, but less so than Pearson's coefficient (5). Kendall's τ can also be used to derive partial correlation coefficients, whereas Spearman's ρ cannot (12). From a computational viewpoint, Spearman's ρ can be calculated in any statistical program that can calculate Pearson's product moment correlation by simply calculating Pearson's correlation on the rank transformed data. Kendall's τ, on the other hand, is more difficult to calculate and usually requires a simple macro or some spreadsheet calculations to count the number of concordant pairs and to solve equation 3.

There are at least 25 different ecological indices alone for characterizing the property of occurrence (6). One reason for choosing the Jaccard index over other indices is that the Jaccard index does not consider disease-free sampling units ("double zero") in its calculation as an indication of association. Other indices, such as the Ochiai and Dice indices also do not use double zero sampling units in their calculation (7,9). However, the Jaccard is simpler to interpret. The Jaccard index simply represents the probability of encountering a sampling unit with both species (diseases) in the population of sampling units with either disease. Other indices, such as the Dice and Ochiai, use alternative weightings of the proportion of sampling units occupied by both species relative to sampling units occupied by only a single species to measure association, making it difficult to interpret these indices biologically (7).

In selecting the Jaccard index, however, one surrenders the opportunity to perform a simple, parametric test of the hypothesis of independence in lieu of, presumably, a more meaningful measure of association. Counting double-zero sampling units would allow the analyst to apply a simple-to-calculate chi-square test designed for two-way tables to test the null hypothesis of independence because the sampling distribution of this statistic is known. In the procedure described above, randomizations were used to derive an empirical, conditional sampling distribution for J and the jackknife was used to derive J's standard error.

The randomization algorithm, although computationally intensive, is relatively easy to program in many spreadsheet or statistical software packages. One result of using randomizations is that the results are conditioned on the underlying properties of the data, which may or may not be well defined. For example, in most standard parametric procedures, independence among observations is a basic assumption. This assumption is not a requirement in order to apply randomizations. The fault, of course, is that a different sampling distribution is generated for every data set, although they are typically similar. Indeed, randomizations are best applied when several independent data sets are available for analysis.

Lastly, although the jackknife was used to derive the standard error of J, the standard error could have just as easily been calculated using the bootstrap. Statisticians tell us that, in general, for a linear statistic there is no loss of information using the jackknife over the bootstrap; the bootstrap is preferred for nonlinear statistics. One disadvantage of the bootstrap is that two different people using the same data will not get the same bootstrap estimate of the standard deviation or confidence interval. This is not the case for the jackknife. It should be noted that the bootstrap cannot be applied indiscriminately, so interested readers should consult more detailed descriptions of the procedure before applying it in their own analyses (2,5).

The analysis of interspecific associations is just one example where nonparametric procedures are well suited. Interested readers who wish to explore the rich set of nonparametric tools available for other analyses are encouraged to consult textbooks (1,3,5,8,12) or review articles (4,11) on this topic.

LITERATURE CITED

1. Dixon, P. M. 1993. The bootstrap and the jackknife: Describing the precision of ecological indices. Pages 290-318 in: Design and Analysis of Ecological Experiments. S. M. Scheiner and J. Gurevitch, eds. Chapman & Hall, New York.
2. Efron, B., and Tibshirani, R. J. 1993. An Introduction to the Bootstrap. Chapman & Hall, New York.
3. Gibbons, J. D. 1985. Nonparametric Methods for Quantitative Analysis. 2nd ed. American Sciences Press, Columbus, OH.
4. Hall, P. 2001. *Biometrika* centenary: Nonparametrics. Biometrika 88: 143-165.

5. Hollander, M., and Wolfe, D. A. 1999. Nonparametric Statistical Methods. 2nd ed. John Wiley & Sons, New York.
6. Legendre, L., and Legendre, P. 1983. Numerical Ecology. Elsevier, New York.
7. Ludwig, J. A., and Reynolds, J. F. 1988. Statistical Ecology. John Wiley & Sons, New York.
8. Manly, B. F. J. 1997. Randomization, Bootstrap and Monte Carlo Methods in Biology. 2nd ed. Chapman & Hall, London.
9. Nelson, S., and Campbell, C. L. 1992. Incidence and pattern of association of pathogens in a leaf spot disease complex on white clover in the Piedmont region of North Carolina. Phytopathology 82:1013-1021.
10. Pethybridge, S. J., and Turechek, W. W. 2003. Analysis of the association among three viruses infecting hop in Australia. Plant Pathol. 52:158-167.
11. Potvin, C., and Roff, D. A. 1993. Distribution-free and robust statistical methods: Viable alternatives to parametric statistics. Ecology 74:1617-1628.
12. Sprent, P., and Smeeton, N. C. 2001. Applied Nonparametric Statistical Methods. 3rd ed. Chapman & Hall/CRC Texts in Statistical Sciences, Boca Raton, FL.
13. Turechek, W. W., and Madden, L. V. 2000. Analysis of the association between the incidence of two spatially aggregated foliar diseases of strawberry. Phytopathology 90:157-170.
14. Turechek, W. W., and Stevenson, K. L. 1998. Effects of host resistance, temperature, leaf wetness, and leaf age on infection and lesion development of pecan scab. Phytopathology 88:1294-1301.