Symposium

New Applications of Statistical Tools in Plant Pathology

Presented at the 94th Annual Meeting of The American Phytopathological Society
July 29, 2002, Milwaukee, WI

New Applications of Statistical Tools in Plant Pathology

K. A. Garrett, L. V. Madden, G. Hughes, and W. F. Pfender

First author: Department of Plant Pathology, 4024 Throckmorton Plant Sciences Center, Kansas State University, Manhattan 66506; second author: Department of Plant Pathology, Ohio State University, Wooster 44691; third author: School of Biological Sciences, University of Edinburgh, West Mains Road, Edinburgh EH9 3JG, U.K.; and fourth author: U.S. Department of Agriculture-Agricultural Research Service National Forage Seed Production Research Center, 3450 SW Campus Way, Corvallis, OR 97331.
Accepted for publication 15 April 2004.

ABSTRACT

Garrett, K. A., Madden, L. V., Hughes, G., and Pfender, W. F. 2004. New applications of statistical tools in plant pathology. Phytopathology 94:999-1003.

The series of papers introduced by this one address a range of statistical applications in plant pathology, including survival analysis, nonparametric analysis of disease associations, multivariate analyses, neural networks, meta-analysis, and Bayesian statistics. Here we present an overview of additional applications of statistics in plant pathology. An analysis of variance based on the assumption of normally distributed responses with equal variances has been a standard approach in biology for decades. Advances in statistical theory and computation now make it convenient to appropriately deal with discrete responses using generalized linear models, with adjustments for overdispersion as needed. New nonparametric approaches are available for analysis of ordinal data such as disease ratings. Many experiments require the use of models with fixed and random effects for data analysis. New or expanded computing packages, such as SAS PROC MIXED, coupled with extensive advances in statistical theory, allow for appropriate analyses of normally distributed data using linear mixed models, and discrete data with generalized linear mixed models. Decision theory offers a framework in plant pathology for contexts such as the decision about whether to apply or withhold a treatment. Model selection can be performed using Akaike's information criterion. Plant pathologists studying pathogens at the population level have traditionally been the main consumers of statistical approaches in plant pathology, but new technologies such as microarrays supply estimates of gene expression for thousands of genes simultaneously and present challenges for statistical analysis. Applications to the study of the landscape of the field and of the genome share the risk of pseudoreplication, the problem of determining the appropriate scale of the experimental unit and of obtaining sufficient replication at that scale.

The disciplines of plant pathology and statistics continue to develop, offering new opportunities for the application of statistics in the biological sciences and new demands for statistical approaches in plant pathology. This series of papers, including this introductory article, offers several perspectives on the contributions of new statistical theory and newly available statistical programs. The papers introduce and highlight statistical methods that are relatively little used in phytopathological research at present, but that have potential for improving the analysis of data from many types of experiments. Taken collectively, the experimental situations appropriate for the statistical tools presented are quite common in plant pathology. Each of the contributed papers provides the context and vocabulary to enable readers to evaluate the utility of the tool to their research, and the references for further exploration. Examples from plant pathology are used to illustrate the analyses, and caveats about inappropriate applications are given.

It is common in plant pathology to estimate the relationship between disease responses and a number of environmental and other predictor variables. Sanogo and Yang (40) provide an overview of multivariate analysis techniques, and provide details about several selected as having greatest utility in plant disease epidemiology research: discriminant analysis, multivariate analysis of variance (ANOVA), correspondence analysis, and canonical correlation analysis.

When little is known about the structural form of complex relationships between response variables and large sets of predictor variables, artificial neural networks can be developed to extract patterns. Francl (13) addresses the history, terminology, and common misconceptions about artificial neural networks, and provides recommendations for their appropriate use in plant pathology research. He discusses applications of neural networks to leaf wetness estimation and infection period prediction in wheat disease forecasting.

Meta-analysis approaches can be used to formally synthesize the results from multiple related studies. Rosenberg et al. (39) describe the potential for meta-analysis in plant pathology and give an example of a meta-analysis of the response of yield to disease severity based on papers published in *Fungicide and Nematicide Tests*. They reference additional, more complex methods of meta-analysis for those interested. They also provide suggestions for authors of research papers to make data more amenable to possible future use in meta-analyses.

While common parametric approaches, such as typical ANOVAs and regression analyses, are well known and convenient, their assumptions are not always met in contexts studied by plant pathologists. Nonparametric approaches are often more appropriate to

Corresponding author: K. A. Garrett; E-mail address: kgarrett@ksu.edu

Publication no. P-2004-0719-01O

such situations. Turechek (48) notes also that nonparametric tools have unique uses where they are superior to parametric tests, or where no parametric approach exists. He discusses nonparametric tests of interspecific association, illustrated with two types of data: co-occurrence of more than one plant disease in the same habitat, and covariance of disease intensity for multiple diseases. An innovative nonparametric procedure is described for testing the significance of the value of the Jacquard index of association based on randomization methodology.

One common topic of study in plant pathology is the time required until an event of interest occurs. For example, plant pathologists frequently study the time that elapses until disease onset or changes in pathogen status. Scherm and Ojiambo (42) note that some popular statistical techniques are inappropriate for such analyses, and they describe the use of survival analysis to extract valid statistical conclusions from the data. They illustrate survival time analyses in a study of the timing of defoliation of blueberry leaves as a function of infection by *Septoria albopunctata*, and reference software for conducting this type of analysis.

Bayesian statistics offer a different framework for statistical analysis by treating the parameter of interest not as a single value but rather as if it were a random variable with a probability distribution. Mila et al. (33) provide the vocabulary and general philosophy of Bayesian statistics, and illustrate its application in genomic analyses, disease mapping, and experimental design.

Here we present an overview of several other general statistical topics relevant to plant pathology: advances in the application of analyses of discrete data using generalized linear models; the analysis of ordinal data; the application of linear mixed models (LMM), including the application of decision theory; model selection; and new applications in the context of microarray analyses. Some of the presentation is an expansion or elaboration of topics mentioned in the other papers.

ANOVA has been the fundamental method used by plant pathologists and other scientists for data analysis and statistical inference for well over a half a century (14). With typical usage, a continuous distribution with a normal distribution has been assumed for the response or dependent variable (Y), and a linear model is fitted to the data using least squares methodology (41). The equations used are known as general linear models. Of course, many response variables of interest to plant pathologists are discrete, such as disease incidence (number or proportion of diseased individuals) or counts of lesions or spores (27,32). The "standard" methodology has been to use a transformation of Y, when possible, that results in a variable that is approximated with a normal distribution. In a sense, this is forcing the data to fit a model that was developed for other purposes, rather than using an appropriate statistical model for the data at hand (18,29). Fortunately, there have been many statistical advances over the last 20 years to allow for a fuller and often more appropriate analysis of data that do not have a normal distribution (6,31,50). These new statistical methods are now being used increasingly by plant pathologists.

Generalized linear models. There are often situations where one cannot assume that models for continuous data are appropriate for discrete data. This would be the case when the number of individuals observed for determining proportions is small in each replicate, or the counts do not have a wide range of values in the particular study. For data of this type, generalized linear models (GLMs) are required (9,18,30). Here, a function of the mean, or expected value, of Y is modeled as a linear function of the variables or factors of interest. This function can be written as $g(\mu)$, where μ is the expectation of Y [$\mu = E(Y)$], and is known as the link function. This is quite different from the regular normal-distribution-based approach of transforming Y to produce $g(Y)$ and then fitting a model to $g(Y)$. In this latter case, a mean of the *transformed* response $E[g(Y)]$ is obtained (say, at each treatment tested), which is not the same as a function of the mean $g[E(Y)] = g(\mu)$ obtained with GLMs. In other words, using GLMs allows means to be calculated correctly using whatever scale is desired rather than forcing researchers to draw inference about, for example, the means of arcsine-transformed responses simply to meet assumptions of normality. With GLMs, it is relatively easy to obtain direct estimates of μ through use of the "inverse link function."

Fitting GLMs to data generally is done using maximum likelihood, a method based on finding parameter estimates that result in the highest probability of observing the actual data obtained. With GLMs, it is straightforward to account for the properties of data from discrete distributions such as the Poisson and binomial, which are appropriate theoretical distributions to consider (at least initially) for counts and proportions, respectively (1,9,18). Among other things, the dependence of the variance of Y on the magnitude of Y is accommodated easily. In contrast, it is more problematic to account satisfactorily for the properties of discrete data using normal-distribution-based methods (41). In particular, one transformation might be best for obtaining a constant variance, but another transformation might be best for a linear scale. Only one of these two transformations can be used, so all the statistical requirements cannot be met with linear models. With GLMs, one often chooses the logit-link function for proportion data and the log-link for counts (those with no upper bound), although other choices are possible (49). For proportions, the analysis is commonly known as logistic regression, especially when the predictor variables of interest are continuous. However, the approach is useful for designed and observational studies where one is relating qualitative factors (e.g., fungicide treatment and cultivar) and quantitative factors (e.g., soil temperature) to responses (11,19,34).

One common problem encountered with GLMs of plant disease data is that the observed variability is greater than that predicted by the binomial or Poisson distributions (9). This is known as overdispersion. Important developments in statistics have shown how to account for overdispersion in GLMs (4,9,41). One approach uses so-called maximum quasi-likelihood rather than maximum likelihood, which essentially rescales the theoretical variability upward to match the observed variability. Another approach involves the use of discrete distributions for overdispersed discrete data. The negative binomial is the most relevant for unbounded counts, and the beta-binomial and logistic-normal-binomial are the most relevant for disease incidence (19,27). Thanks to the development of specialized software and procedures with commercial programs such as SAS and EGRET (Cytel Software Corp., Cambridge, MA), it is now easier to use these more complicated models in linear models for data analysis (18–20,26,37). Furthermore, GLM-based analysis of disease incidence and lesion counts from observational (survey) studies can be of direct benefit in developing efficient sampling protocols for either estimating mean disease levels or testing hypotheses about mean level (16,17,28).

Ordinal data. Plant pathologists sometimes measure disease intensity using an ordinal scale, such as 1 for no symptoms, 2 for mild symptoms, 3 for major disease symptoms, and 4 for a dead plant. This produces discrete data, and the random variable is of the ordinal categorical type (41). With sufficient observations, one can analyze these data with GLMs (pages 379–383 in literature citation 41) by assuming a multinomial distribution for the variable. This approach becomes challenging when both overdispersion and small numbers of observations (e.g., when there are not several values of each ordinal category) occur (2). A very useful alternative is to use nonparametric analysis (45). Turechek (48) in this symposium discusses several features of nonparametric analysis. The nonparametric approach has been well known and advocated for many years, and there are many programs that can perform such analyses. However, a limitation of the nonparametric approach was that, until the last 5 years or so,

there had been no satisfactory theoretical foundation for modeling data originating in factorial designs, including split plots and repeated measures (43). Fortunately, thanks to contributions by Brunner et al. (5) and Brunner and Puri (6), there is now a relatively straightforward statistical approach for handling these designs. Although parametric models offer the most flexibility, generality, and statistical power for many possible experimental designs, when the properties of the response variable (*Y*) justify their use, the new nonparametric methods are highly advantageous for data that pose problems for parametric analysis. Details are available in Shah and Madden (43).

Linear mixed models. Terms in linear models fitted to data represent the effects of variables (or factors) either controlled by the researcher in planned experiments or simply measured in observational studies, plus a residual error that represents all the variability in *Y* (or transformation of *Y*) not accounted for by the other terms. It is long recognized that terms in models can represent fixed or random effects on the response variable. A fixed-effects variable is one at which the levels in the study represent all possible levels, or all the levels of interest to the investigator (38, 41). Examples would be fungicide treatment, biocontrol agent, inoculum dose, and so on. In contrast, a random-effects variable (or factor) is one for which the levels in the study represent only a random sample of a larger set of potential levels, or for which one is not interested in the specific results for each level in the study (24) but wishes to draw inference about the variability of a larger population. Examples could be environment (block or location) and host crop genotype. To clarify, genotype could be either of the random- or fixed-effects type, depending on whether one is interested in the population of all genotypes (e.g., in a population genetics study) or in the specific cultivars studied (e.g., for deciding if a new cultivar is disease resistant). In other words, the type of effect depends, in part, on the objectives of the research.

The residual error in a model is a random-effects term. Most variables or factors studied by plant pathologists are of the fixed-effects type. However, certain aspects of experimental design and data collection create random-effects terms for models (41). Any clustering of the data, which occurs with blocking, sampling and subsampling within experimental units, and collection of data over time on the same experimental units (repeated measures) create random-effects terms in the model that need to be fitted to data. For example, for a split plot design, one needs both the residual error term and a random-effects term for the variation of *Y* among the whole–plot experimental units. Linear models with more than one random effect (residual and at least one more random effect term) and with fixed effects are known as LMMs (25). The importance of mixed-effects models for analysis of data from many types of experiments of relevance to plant pathologists has been understood for many decades (31,50). However, until the last 20 years, true mixed-effects modeling was very difficult, except for the specialist statistician. Thanks to major advances in statistical theory and methodology, with concomitant advances in computer algorithms and increasing memory and speed of computers, it is now possible for experimental scientists to fit LMMs to data and correctly interpret estimated parameters (such as the effect of a treatment on mean *Y*) (4).

The biological literature is full of papers claiming to have used a LMM model for data analysis. However, until recently, these models were not truly of the LMM type. Many computer programs, such as PROC GLM of SAS (SAS Institute, Cary, NC), actually fit pure fixed-effects models to data, even when one designates certain terms as random effects (25). After fitting the model, the program estimates the appropriate variance components and then conducts *F* tests for significance. For completely balanced designs, these *F* tests are correct with these ad hoc post-model fitting calculations. However, standard errors of certain means can be incorrect even with a balanced situation (25). Moreover, a very slight imbalance, such as just one missing value, can lead to misleading *F* statistics for hypothesis testing and incorrect standard errors of means and differences of means! Obviously, this can greatly affect inference.

Fitting a true LMM to data is a computer-intensive iterative process and is done typically using maximum likelihood estimated (MLE) or restricted maximum likelihood estimation (REML) methods (24). Programs such as PROC MIXED in SAS perform true LMM fitting, rather than the unsatisfactory traditional approach explained in the previous paragraph. Slowly, plant pathologists and other scientists are migrating from the use of programs such as GLM of SAS to MIXED. The advantages are clear, as discussed in the SAS/SAT manual (SAS Institute) and numerous other references (4,24,25,31,38). In addition to producing correct test statistics and standard errors, the LMM approach allows for direct incorporation of many types of data and design features, such as unequal variances (variance heterogeneity [51], a property of many disease data [26,27,29]), temporal correlation of data within experimental units, and spatial correlation of *Y* across experimental units (41).

Use of LMMs does involve somewhat of a mind shift in terms of data analysis. For instance, with the traditional teaching of ANOVA for fixed- and mixed-effects models (really fixed effects with post-model fitting calculations), researchers learned about interpretation of factor effects in terms of reductions in sums of squares, and chose proper *F* tests based on expected mean squares for terms in the models. With true LMMs, however, there are no sums of squares or mean squares (24)! Instead, there are log-likelihoods and likelihood-ratio statistics, and the use of "Wald statistics" for the calculation of *F* tests. Moreover, degrees of freedom for tests can be very different from what would have been determined from now out-of-date use of fixed-effects models for data that truly require use of a LMM.

Generalized linear mixed models. Just as GLMs can be expanded to handle overdispersion, they can be expanded to handle more than just one random effect (4,9,37). In particular, generalized linear mixed models (GLMMs) can be used for discrete data such as disease incidence and counts of lesions to account for multiple random effects (e.g., location effect, variation within- and between-experimental units, and sampling variation). Some of the details are given in Madden et al. (29) for a disease incidence response variable. GLMMs are considerably more complicated than GLMs or LMMs, and all of the statistical properties of GLMMs are not yet resolved. Nevertheless, this approach nicely complements the LMMs used for continuous data.

As indicated by Littell (24), there are many challenges in making the transition from linear models to LMMs for data analysis. There may be even more challenges to moving from GLMs to GLMMs (chapter 8 in literature citation 41). Nevertheless, the statistical evidence is clear that the transition is worth the effort for many experiments or observational studies. Now that textbooks have been written to teach the material to non-statisticians (4,25,41,50), at least to quantitative biologists, and software is available in programs such as SAS and GENSTAT, LMMs and GLMMs can be utilized by biologists. The modeling efforts are more involved, however, and plant pathologists are advised to work with consulting statisticians on the analysis whenever possible.

Decision theory. The application of decision theory provides a framework for evidence-based decision-making in plant pathology. In the clinical context, Ashby and Smith (3) wrote "Evidence-based medicine requires an integrated assessment of the available evidence, and associated uncertainty, but there is also an emphasis on decision-making, for individual patients, or at other points in the health-care system." From a phytopathological perspective, we could say that evidence-based disease management requires an integrated assessment of the available evidence, and associated uncertainty, with an emphasis on decision-making. The decision maker will often be the individual farmer or grower (or

an advisor), but sometimes decisions are made at other points in the agricultural system, for example by buyers of crops grown on contract or by regulatory authorities. A short list of questions summarizes the disease management decision-making process, as follows:

- Who is the decision maker?
- What are the risk factors?
- What are the possible actions?
- What are the consequences?
- What are the utilities (such as costs and benefits to the decision maker)?

We need to identify the decision maker because an individual farmer, for example, might reach a different decision about a disease management problem than a regulatory authority. For evidence-based decision-making we require data on the relevant risk factors. In the context of disease management, risk factors relating to the pathogen, the host, and the environment are combined into a statistical prediction rule, or predictor. In the simplest case, which nevertheless covers many important scenarios (47), there are just two actions to consider: for example, apply/withhold treatment. However, there is uncertainty attached to the consequences of the possible actions because we do not know whether a crop actually requires treatment or not. If we waited until the end of the season to find out for sure that a crop needed treatment, it would be too late to do anything about it. So instead we act on the basis of a predictor. Uncertainty arises from the imperfection of the predictor. Sensitivity (the proportion of positive predictions that are correct) and specificity (the proportion of negative decisions that are correct) characterize the accuracy of a predictor and have important applications in evidence-based decision-making (53). An assessment of utilities requires information on costs and benefits. These may vary according to the point of view adopted by the decision maker. Yuen (53) and Yuen and Hughes (54) discuss a method to combine sensitivity and specificity with information on the prior probability of disease, using Bayes' theorem, to calculate the posterior probability of disease, given the evidence related to risk factors. Although disease management has not been explored traditionally from a formal Bayesian perspective, many of the decisions that must be made for disease control are considered easily in this context. Analyses of this type may ultimately lead to an improved or more efficient disease control decision-making process. Mila et al. (33), in this symposium, offer a more thorough presentation on Bayesian analysis. Brown and Prescott (4) describe how to conduct LMM analysis in a Bayesian manner.

Model selection. While biologists have traditionally stressed hypothesis testing as a statistical approach, emphasis has shifted in recent years. For example, writing for an ecological audience, Hilborn and Mangel (15) emphasized the use of likelihood and Bayesian methods in contrast to testing null hypotheses. Burnham and Anderson (7) have made an important contribution to changes in the fundamental orientation of ecologists toward use of statistics. As a reviewer of the first edition of their book wrote (12), "Abandon all *P*-values, ye who enter here!" Burnham and Anderson (7) instead take an information-theoretic approach by specifying a set of plausible models a priori and then selecting among models using measures such as Akaike's information criterion (AIC) (41). This criterion can be applied to choose the model that summarizes the data most efficiently. In fact, the AIC is a critical tool in LMM analysis of data from designed and observational experiments (25).

Microarray analysis. Until recently, complex statistical analyses in plant pathology have been applied most commonly in population and community level studies. In these contexts, the variance may be high but it may also be possible to include a large number of replicates to increase statistical power. New technologies in the study of gene expression have created the need for new statistical methods, as well as new applications of old methods, in order to handle the large volumes of data produced and practical limitations on replication. Microarray technologies now make it possible for researchers to simultaneously analyze gene expression in thousands of genes per microarray from a single experimental unit, but the cost of analysis for each experimental unit often severely limits the number of replicates. Microarray analysis of gene expression raises three familiar issues for statistical applications (23,36,44). First, experiments must be designed strategically because of the cost of replication. While some microarrays, such as those produced by Affymetrix (Santa Clara, CA), are designed to accommodate a single sample, two (or more) samples, such as a comparison between treatment and control, can be processed in each cDNA microarray with the samples distinguished by color. For more complicated treatment structures, clever incomplete block designs (41) may be needed, where each microarray is treated as an incomplete block (8); microarray analysts often refer to variations on these designs as "loop" designs. Second, the hundreds or thousands of genes studied simultaneously mean that hundreds or thousands of estimates of up- or down-regulation are tested to determine whether there is evidence for a change due to treatment. In an effort to control the experiment-wise type I error rate, the probability of a false positive over the whole experiment, researchers have applied approaches such as simple or modified Bonferroni corrections or nonparametric resampling to adjust the error rate. For a simple Bonferroni correction, the standard for statistical significance becomes very strict, making it difficult to detect effects that are real. False discovery rate approaches offer an alternative by considering the probability that responses identified as positive are incorrectly identified (46 and software referenced within). Third, clustering methods are applied to group genes with similar responses to experimental treatments.

The evolution of statistical approaches to microarray analysis presents an interesting example of how the need for statistical techniques can emerge in one field with little awareness of the techniques already available and practiced in other fields. Because of the great expense associated with microarray processing, studies in the early stages of its development have included little or even no replication. Instead, the significance of up- or down-regulation observed for any given gene has been evaluated simply based on the estimated size of the change, using arbitrary cut-offs for significance such as a twofold difference. As replication has become more feasible economically in microarray experiments, there is still a tendency to evaluate results in terms of "fold change." Concurrently, some statisticians have developed applications of mainstream statistical techniques to this study area. For example, Wolfinger et al. (52) present a method for modeling gene expression responses from cDNA microarray experiments using two interconnected LMMs. Jin et al. (22) give an example of such an application. Mila et al. (33) also discuss a Bayesian approach to microarray analysis in this symposium series.

Pseudoreplication in large-scale ecological studies and in microarray analyses. Pseudoreplication was a significant topic for discussion among ecologists when Hurlbert's monograph detailing the phenomenon (21) was published. Hurlbert defined pseudoreplication as "the use of inferential statistics to test for treatment effects with data from experiments where either treatments are not replicated (though samples may be) or replicates are not statistically independent." For example, multiple samples from the same experimental unit, such as an individual organism or an individual experimental plot, could be pseudoreplicates. Microarray analysis is a new discipline in which pseudoreplication is a risk if samples are drawn only from the same individual or genetically similar individuals rather than drawing from an appropriately diverse population. A backlash arose among ecologists in defense of approaches that could be conceptualized as pseudoreplication but that were unavoidable because of the large spatial scale of inquiry (10,35). Also, with the new availabil-

ity of more flexible statistical computing packages, LMMs can be used to account for pseudoreplication by explicitly determining the correlations of the response variable among all individuals, rather than requiring the assumption that observations are independent (41). Because the landscape of the field and the landscape of the genome offer similar challenges for appropriate statistical analyses, both areas stand to benefit from statistical solutions developed in either.

ACKNOWLEDGMENTS

Thanks to the participants in this symposium for all their contributions and to the APS Epidemiology Committee for its sponsorship. We appreciate the critical review of this manuscript by S. E. Travers and *Phytopathology* reviewers. This work was supported in part by NSF grants DEB-0130692, EPS-0236913 with matching funds from the Kansas Technology Enterprise Corporation, and EPS-9874732 with matching support from the State of Kansas, and by USDA Grant 2002-34103-11746. This is Kansas State Experiment Station Contribution No. 04-258-J.

LITERATURE CITED

1. Agresti, A. 2002. Categorical Data Analysis. Wiley-Interscience, New York.
2. Agresti, A., and Natarajan, R. 2001. Modeling clustered ordered categorical data: A survey. Int. Stat. Rev. 69:345-371.
3. Ashby, D., and Smith, A. F. M. 2000. Evidence-based medicine as Bayesian decision-making. Stat. Med. 19:3291-3305.
4. Brown, H., and Prescott, R. 1999. Applied Mixed Models in Medicine. John Wiley & Sons, Chichester, UK.
5. Brunner, E., Domhof, S., and Langer, F. 2002. Nonparametric Analysis of Longitudinal Data in Factorial Experiments. John Wiley & Sons, New York.
6. Brunner, E., and Puri, M. L. 2001. Nonparametric methods in factorial designs. Stat. Papers 42:1-52.
7. Burnham, K. P., and Anderson, D. 2002. Model Selection and Multi-Model Inference. 2nd ed. Springer-Verlag, New York.
8. Churchill, G. A. 2002. Fundamentals of experimental design for cDNA microarrays. Nature Gen. Suppl. 32:490-495.
9. Collett, D. 2002. Modelling Binary Data. 2nd ed. CRC Press, Boca Raton, FL.
10. Cottenie, K., and De Meester, L. 2003. Comment to Oksanen (2001): Reconciling Oksanen (2001) and Hurlbert (1984). Oikos 100:394.
11. De Wolf, E. D., Madden, L. V., and Lipps, P. E. 2003. Risk assessment models for wheat Fusarium head blight epidemics based on within-season weather data. Phytopathology 93:428-435.
12. Ellison, A. M. 1999. Abandon all *P*-values, ye who enter here! Ecology 80:2129-2130.
13. Francl, L. J. 2004. Squeezing the turnip with artificial neural nets. Phytopathology 94:1007-1012.
14. Gilligan, C. A. 1986. Use and misuse of the analysis of variance in plant pathology. Pages 225-261 in: Advances in Plant Pathology, Vol. 5. Academic Press, New York.
15. Hilborn, R., and Mangel, M. 1997. The Ecological Detective: Confronting Models with Data. Princeton University Press, New Jersey.
16. Hughes, G., and Gottwald, T. R. 1998. Survey methods for assessment of citrus tristeza virus incidence. Phytopathology 88:715-723.
17. Hughes, G., and Gottwald, T. R. 1999. Survey methods for assessment of citrus tristeza virus incidence when *Toxoptera citricida* is the predominant vector. Phytopathology 89:487-494.
18. Hughes, G., and Madden, L. V. 1995. Some methods allowing for aggregated patterns of disease incidence in the analysis of data from designed experiments. Plant Pathol. 44:927-943.
19. 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.
20. Hughes, G., and Samita, S. 1998. Analysis of patterns of pineapple mealybug wilt disease in Sri Lanka. Plant Dis. 82:885-890.
21. Hurlbert, S. H. 1984. Pseudoreplication and the design of ecological field experiments. Ecol. Monog. 54:187-211.
22. Jin, W., Riley, R. M., Wolfinger, R. D., White, K. P., Passador-Gurgel, G., and Gibson, G. 2001. The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. Nature Gen. 29:389-395.
23. Knudsen, S. 2002. A Biologist's Guide to Analysis of DNA Microarray Data. Wiley-Interscience, New York.
24. Littell, R. C. 2002. Analysis of unbalanced mixed model data: A case study comparison of ANOVA versus REML/GLS. J. Agric. Biol. Environ. Stat. 7:472-490.
25. Littell, R. C., Milliken, G. A., Stroup, W. W., and Wolfinger, R. D. 1996. SAS System for Mixed Models. SAS Institute, Cary, NC.
26. Madden, L. V., Ellis, M. A., Lalancette, N., Hughes, G., and Wilson, L. L. 2000. Evaluation of a disease warning system for downy mildew of grapes. Plant Dis. 84:549-554.
27. Madden, L. V., and Hughes, G. 1995. Plant disease incidence: Distributions, heterogeneity, and temporal analysis. Annu. Rev. Phytopathol. 33:529-564.
28. Madden, L. V., and Hughes, G. 1999. Sampling for plant disease incidence. Phytopathology 89:1088-1103.
29. 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.
30. McCullagh, P., and Nelder, J. A. 1989. Generalized Linear Models. 2nd ed. Chapman & Hall, London.
31. McCulloch, C. E., and Searle, S. R. 2001. Generalized, Linear, and Mixed Models. John Wiley & Sons, New York.
32. McRoberts, N., Hughes, G., and Madden, L. V. 2003. The theoretical basis and practical application of relationships between different disease intensity measurements in plants. Ann. Appl. Biol. 142:191-211.
33. Mila, A. L., and Carriquiry, A. L. 2004. Bayesian analysis in plant pathology. Phytopathology 94:1027-1030.
34. Mila, A. L., Carriquiry, A. L., and Yang, X. B. 2004. Logistic regression modeling of prevalence of soybean Sclerotinia stem rot in the North-Central region of the United States. Phytopathology 94:102-110.
35. Oksanen, L. 2001. Logic of experiments in ecology: Is pseudoreplication a pseudoissue? Oikos 94:27-38.
36. Parmigiani, G., Garrett, E. S., Irizarry, R. A., and Zeger, S. L., eds. 2003. The Analysis of Gene Expression Data: Methods and Software. Springer-Verlag, New York.
37. Piepho, H.-P. 1999. Analysing disease incidence data from designed experiments by generalized linear mixed models. Plant Pathol. 48:668-674.
38. Piepho, H.-P., Buchse, A., and Emrich, K. 2003. A hitchhiker's guide to the mixed model analysis of randomized experiments. J. Agron. Crop Sci. 189:310-322.
39. Rosenberg, M. S., Garrett, K. A., Su, Z., and Bowden, R. L. 2004. Meta-analysis in plant pathology: Synthesizing research results. Phytopathology 94:1013-1017.
40. Sanogo, S., and Yang, X. B. 2004. Overview of selected multivariate statistical methods and their use in phytopathological research. Phytopathology 94:1004-1006.
41. Schabenberger, O., and Pierce, F. J. 2002. Contemporary Statistical Models for the Plant and Soil Sciences. CRC Press, Boca Raton, FL.
42. Scherm, H., and Ojiambo, P. S. 2004. Applications of survival analysis in botanical epidemiology. Phytopathology 94:1022-1026.
43. Shah, D. A., and Madden, L. V. 2004. Nonparametric analysis of ordinal data in designed factorial experiments. Phytopathology 94:33-43.
44. Speed, T. P., ed. 2003. Statistical Analysis of Gene Expression Microarray Data. Chapman & Hall, Boca Raton, FL.
45. Sprent, P., and Smeeton, N. C. 2001. Applied Nonparametric Statistical Methods. Chapman & Hall, Boca Raton, FL.
46. Storey, J. D., and Tibshirani, R. 2003. Statistical significance for genomewide studies. PNAS 100:9440-9445.
47. Swets, J. A., Dawes, R. M., and Monahan, J. 2000. Better decisions through science. Sci. Am. 283:70-75.
48. Turechek, W. W. 2004. Nonparametric tests in plant disease epidemiology: Characterizing disease associations. Phytopathology 94:1018-1021.
49. Turechek, W. W., and Madden, L. V. 2003. A generalized linear modeling approach for characterizing disease incidence in a spatial hierarchy. Phytopathology 93:458-466.
50. Verbeke, G., and Molenberghs, G., eds. 1997. Linear Mixed Models in Practice; A SAS-Oriented Approach. Lecture Notes in Statistics, No. 126. Springer-Verlag, Berlin.
51. Wolfinger, R. D. 1996. Heterogeneous variance-covariance structures for repeated measures. J. Agric. Biol. Environ. Stat. 1:205-230.
52. Wolfinger, R. D., Gibson, G., Wolfinger, E. D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C., and Paules, R. S. 2001. Assessing gene significance from cDNA microarray expression data via mixed models. J. Comp. Biol. 8:625-637.
53. Yuen, J. 2003. Bayesian approaches to plant disease forecasting. Online. Plant Health Progress doi:10.1094/PHP-2003-1113-06-RV.
54. Yuen, J. E., and Hughes, G. 2002. Bayesian analysis of plant disease prediction. Plant Pathol. 51:407-412.