STATISTICAL METHODOLOGY IN AGRICULTURE AND HORTICULTURE

A. Mead
Warwick HRI, University of Warwick, U.K

Keywords: Variability, experimental design, analysis of variance (ANOVA), regression, generalized linear model (GLM), analysis of deviance, restricted maximum likelihood (REML), spatial data, precision agriculture, on-farm experimentation.

Contents

Summary

Many modern statistical techniques were first developed for use in agricultural research, and many basic statistical tools are still important for such research. Good experimental design, following the basic principles of replication, blocking and randomization, allows the control of anticipated environmental variation and the estimation of treatment effects in the presence of such variation. Analysis of variance provides a wide-ranging approach to the analysis of data from designed experiments, aiding the interpretation of the results of complex experiments. Regression analysis can be used to explore the relationships between a quantitative response variable and one or more quantitative explanatory variables. Linear regression techniques primarily provide an exploratory approach, whilst non-linear regression techniques allow the modeling of responses using biologically realistic relationships. Generalized linear models (GLMs) provide an important tool for working with the non-Normally distributed data that is common in the crop protection experimentation that frequently occurs in agricultural horticultural research, with log-linear models (for count data) and probit or logit models (for counts as proportions) being important specific cases. Residual (or restricted) maximum likelihood (REML) is a relatively recent addition to the agricultural statistician's toolbox, providing an approach to the analysis of linear mixed models, such as

unbalanced experimental designs, and the estimation of components of variance. Future developments of statistical methodology will be important in three areas of agricultural research – the analysis of spatial data, the development of precision agriculture techniques, and on-farm experimentation.

1. Introduction

The use of statistical techniques in agriculture goes back many years, and, in fact, many of the modern statistical techniques were first developed for use in agricultural research. Early developments, due to R.A. Fisher at Rothamsted Experimental Station in the U.K. in the 1920s, included the basic principles of experimental design – replication, randomization and blocking – and the analysis of variance, and these techniques are still the basic tools used in agricultural research today. These techniques, in common with many statistical methods, were developed to cope with the inherent variability associated with experimentation using biological material. In fact, it is the need to explain or allow for the extensive variation often found in experimental biological data that has driven, and still drives, the development of statistical techniques. By using the correct statistical tools we can separate the signal from the noise within our data – if we do not handle the experimental variability properly we run the danger of being unable to draw any useful conclusions from our data.

In the design of experiments, the grouping or blocking of experimental units can be used to eliminate the effects of systematic changes in environmental conditions (the experimental units within a block are assumed to be as similar as possible). The randomization of treatments to units can protect against unknown variability. Replication provides the basis for the comparison of treatments, allowing the assessment of whether the differences between treatments are large relative to the variation between replicate observations on each treatment. The most commonly used experimental design is the "randomized complete block design", with a complete replicate of the set of treatments appearing in each block of experimental units, but many more complex experimental designs have been developed, based on Fisher's principles, to cope with more complex experimental conditions. These include incomplete block designs, row-and-column designs (e.g. Latin squares) and split-plot designs. The analysis of variance technique separates the variation in observed results into that due to the applied treatments and that due to the experimental environment, and hence allows the assessment of whether observed treatment differences are important relative to the underlying variation between experimental units. The analysis of variance technique for analyzing data from designed experiments is readily available in most statistical computing packages

Where applied treatments are quantitative, it is often of more interest to determine the form of relationship between the response variable and these explanatory variables using regression analysis. Simple linear regression is concerned with fitting the simplest of relationships, a straight line, between the response variable and a single explanatory variable, with the parameters of the line (slope, intercept) determined to minimize the variance in the response variable about the fitted line. It is important to realize that the adjective *linear* in *simple linear regression* refers not to the fitting of a straight line, but to the relationship between the response variable and parameters being linear.

Extensions of this linear regression approach include multiple linear regression (more than one explanatory variable), linear regression with groups (including a qualitative treatment factor and allowing parameters to vary with different levels of this factor) and polynomial regression (quadratic, cubic, … relationships between response variable and explanatory variable). Many real biological relationships, however, are not well described by the range of models that can be constructed within the linear regression framework, but require the use of models where the response variable is related to the parameters in a non-linear fashion. Advances in computing power now make the fitting of such non-linear regression models relatively simple, and many standard non-linear response functions are readily available in most statistical computing packages. These include models based on the exponential function (for example, to describe the decay of pesticides in soil or unconstrained growth), sigmoid functions, such as the logistic and Gompertz curves (to describe constrained growth or for dose-response studies), and rational functions, including inverse polynomials (used to describe the relationship between crop yield and applied nutrient levels).

The analysis of variance and regression analysis methods that are mentioned above have an underlying assumption that the response variable is continuous and Normally distributed. However, much of the data collected in agricultural and horticultural research, particularly in relation to crop protection research, are in the forms of discrete counts (numbers of weeds, insects, disease lesions) or proportions based on counts (numbers of diseased fruit per tree, or of insects killed by some treatment), and therefore do not satisfy these assumptions. For example, count data may follow a Poisson distribution and proportions based on counts may follow a Binomial distribution. In this situation two approaches are possible – to find some transformation of the data that allows this assumption to be satisfied or to use an alternative form of analysis that takes account of the distributional form of the data. The development of Generalized Linear Models (GLMs) by McCullagh, Wedderburn and Nelder provided a solution to the latter approach, allowing the analysis of data for a range of non-normal distributions, within the same basic structure as for analysis of variance and regression analysis. Of particular interest within this framework are log-linear models for count data and probit/logit models for proportions based on counts, these latter approaches being particularly important for the analysis of bioassay experiments.

Another, more recent, development is the method of residual (or restricted) maximum likelihood (REML). This algorithm estimates the treatment effects and variance components in a linear mixed model, that is a linear model with both fixed and random effects. The standard model used in an analysis of variance of a designed experiment is of this form, with, in general, the applied treatment factors being fixed effects and the blocking factors being random components. One use of the REML method, however, is as an alternative to regression analysis for the analysis of data from unbalanced experimental designs, where analysis of variance cannot be used. The advantage of the REML approach over regression is that it can account for more than one source of variation in the data (e.g. an unbalanced row-and-column design), and provide estimates of the variance components associated with each of the random (blocking) terms in the model. The REML approach can also be used to obtain information on the sources and sizes of variability in data sets, for example to assess the relative sizes of different sources of variability, or to combine information over similar experiments conducted at

different sites or times, to obtain treatment estimates that make use of the information from all the experiments.

There are a number of areas where future development of statistical methodology will be important in agriculture and horticulture. One is the analysis of spatial data. Whilst spatial statistical methods have been developed and used for many years, particularly geostatistical methods in the mining industry and hydrology, there has been relatively little use of such methods in agriculture and horticulture. Interest in the spatial distributions of plants, pests, diseases, nutrients and pesticides, however, is now becoming important both in understanding the biological processes behind agricultural production, and particularly in the development of precision agriculture approaches to apply, for example, pesticides or fertilizers to match the requirements of small areas of crop. Another area where development of statistical methodology is needed is for on-farm experimentation, involving the assessment of experimental methods when scaled-up from small experimental plots to whole field (or even whole farm) experiments.

-
-
-

Bibliography

Cochran, W.G. & Cox, G.M. (1957). *Experimental Designs (Second edition)*, 611 pp. New York: Wiley. [A comprehensive description of a wide range of experimental designs and their analysis]

Cressie, N.A.C. (1993). *Statistics for Spatial Data*, 900 pp. New York, Wiley. [A wide-ranging description of approaches to the analysis of spatial data]

Dobson, A.J. (2001). *An Introduction to Generalised Linear Models (Second edition)*, 240 pp. London: Chapman & Hall. [A gentle and readable introduction to generalized linear models]

Draper, N.R. & Smith H. (1998). *Applied Regression Analysis (Third edition)*, 736 pp. New York: Wiley. [A comprehensive guide to the application of different regression approaches to real data]

Finney, D.J. (1971). *Probit Analysis (Third Edition)*. 333 pp. Cambridge: Cambridge University Press. [The original text on probit analysis]

McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models (Second edition)*. 511 pp. London: Chapman & Hall. [The comprehensive text on the development of generalised linear models]

Mead, R., Curnow, R.N. & Hasted, A.M. (2002). *Statistical Methods in Agriculture and Experimental Biology (Third edition)*. 472 pp. London: Chapman & Hall. [A general statistical text providing a wide range of statistical methods, written with the non-statistician in mind]

Mead, R. (1988). *The Design of Experiments: Statistical Principles for Practical Application*, 620 pp. Cambridge: Cambridge University Press. [A more recent approach to the design of real experiments]

Snedecor, G.W. & Cochran, W.G. (1989). *Statistical Methods (Eighth edition).* 503 pp. Ames: Iowa State University Press. [A general statistical text providing details of a wide range of statistical methods]

Sokal, R.R. & Rohlf, F.J. (1995) *Biometry (Third edition).* 887 pp. New York: Freeman. [A general statistical text providing details of a wide range of statistical methods, possibly more accessible to non-statisticians]

Biographical Sketch

Andrew Mead studied Statistics at the University of Bath (B.Sc. 1986), and then completed the M.Sc. in Biometry (including a dissertation on "Multidimensional Scaling and its application in Sensory Analysis") at the University of Reading (1987). He joined the Biometrics group at Horticulture Research International now Warwick HRI, University of Warwick at Wellesbourne, near Warwick, in the UK in 1987. Warwick HRI is the principal UK organization tasked with carrying out horticultural research and development (R&D) and transferring the results to industry.

He was admitted to the status of Chartered Statistician (C.Stat.) by the Royal Statistical Society in 1993, and served on the British Regional Committee of the International Biometric Society from November 1995 to November 1998, and from November 1999 to November 2000. He was elected as British Regional Secretary of the International Biometric Society in November 2000, and elected to the International Council of the International Biometric Society in April 2002. He is currently co-Chair of the IBS Strategic Plan Committee. He is also a member of the European Weed Research Society, and acts as Statistical Consultant to the editorial board of *Weed Research*. He was appointed to the External Advisory Panel for the M.Sc. in Biometry at the University of Reading in July 2002 and to the Advisory Committee for M.Sc. in Statistics at the University of Sheffield in October 2004.

Areas of expertise include experimental design for field, glasshouse and mushroom trials and laboratory studies, statistical analysis of experimental data, including modeling and multivariate approaches, and the provision of statistical and mathematical training for biologists. Andrew is responsible for leading the Warwick HRI Biometrics Training Programme, providing appropriate statistical and mathematical training for both PhD students and staff within Warwick HRI.