Andhra Pradesh J Agril. Sci : 1(3): 8-10, 2015

STATISTICAL DESIGNS AND REGRESSION ANALYSIS IN AGRICULTURAL SCIENCES

G. MOHAN NAIDU* AND P. SUMATHI

Department of Statistics and Maths, S.V. Agricultural College, Tirupati – 517 502

ABSTRACT

Statistics is important in the field of agriculture, because it provides tools to analyze collected data. 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 experimental designs, allows the control of anticipated environmental variation and the estimation of treatment effects in the presence of such variation. ANVOA 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.

KEYWORDS: ANOVA, Experimental design, GLM, Hypothesis testing, Regression, p-value, Variability

INTRODUCTION

Statistical education for agriculturists tries to give them a solid foundation in statistics. A wide use of statistical methods in order to allow the students to apply these techniques in many fields of agricultural sciences like filed crop production, livestock, veterinary medicine, agricultural mechanization, water resources, agricultural economics and other fields. 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 United Kingdom in 1920s included the basic principles of experimental design – replication, randomization and local control – and the analysis of variance (ANOVA), and 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.

DESIGN OF EXPERIMENTS AND ANALYSIS OF VARIANCE

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 conditions. These include incomplete block designs, row-and-column designs (e.g. Latin squares) and spilt 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 ANOVA technique for analyzing data from designed experiments is readily available in most statistical computing packages.

*Corresponding author, E-mail: naidu_svag2001@yahoo.com

REGRESSION ANALYSIS

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… etc., relationships between response variable and explanatory variables). 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 computer 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 ANOVA 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 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 data. The development of General Linear Models (GLMs) by McCullagh 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 frame work 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.

There are a number of areas where future development of statistical methodology will be important in agriculture. One is the analysis of spatial data. Whilst spatial statistical methods have been developed and used for many years, particularly geo-statistical methods in the mining industry and hydrology, there has been relatively little use of such methods in agriculture. 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.

STATISTICAL ERROR IN HYPOTHESIS TESTING

There are two types of error or incorrect conclusions possible in hypothesis testing and possibilities in which the statistical test falsely indicates that significant differences exists between the two or more groups and also analogously to a wrong positive results. Rejection of null hypothesis (H_0) when it is true is called Type-I error and acceptance of null hypothesis (H_0) when it is false and it is known as Type-II error and Type-II error is more harmful than Type-I error (Keppel, 1978; Gupta and Kapoor, 1970).

The probability of Type-I error is known as level of significance (á) and the probability of type II error is known as the power of the test â or (1-á) (Keppel, 1978; Gupta and Kapoor, 1970). By convention, statistical significance is generally accepted if the probability of making Type-I error is less than 0.05, which is commonly denoted as p<0.05 (Elenbaas *et al.,* 1983). The probability of Type-II error is more difficulty to derive than probability of type-I error, actually it is not one single probability value. The probability of type-II error (â) is often ignored by researcher (Freeman *et al.,* 1978). The probability of type- I error (á) and probability of type-II error (â) are inter-related. As á arbitrarily decreased, â is increased. Similarly, á is increased, â is decreased (Hopkins and Glass, 1978; Keppel, 1978).

P-VALUE

The p value is the probability to observe effects as big as those seen in the study if there is really no difference between the groups or treatments. The reasoning of hypothesis testing and p values is convoluted. The p values helps to answering whether this apparent effect is likely to be actual or could just by chance or sampling fluctuation. The p values give the magnitude of difference present between populations. In calculation of p values, first assume that no true difference between the two groups/treatments. The p values allow the assessment of findings that are significantly different or not. If the p value is small, the findings are unlikely to have arisen by chance or sampling fluctuations, reject the null hypothesis. If the p is large, the observed difference is plausibly chance finding, we do not reject the null hypothesis. By convention, p value of less than 5 per cent is considered small or significant. Sometimes p value is less than 1 per cent or 0.01, called as highly significant (Gupta and Kapoor, 1970; Rao, 1985).

CONCLUSIONS

Many modern statistical techniques were first developed for the use in agricultural research, and many basic statistical tools are still important for such research. Good experimental design, following the basic principles of replication, randomization and local control, allows the control of anticipated environmental variation and the estimation of treatment effects in the presence of such variation. ANVOA 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 explanatory approach, whilst non-linear regression techniques allow the modeling of responses using biologically realistic relationships. Generalized linear models (GLM) 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 research, with log-linear models (for count data) and probit or logit models (for counts as proportions) being important specific cases. 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. In this paper, the role of statistical research design and regression application of basic techniques in agricultural research, have been emphasized scientifically.

REFERENCES

Elenbass R.M., J.K. Elenbass and P.G. Cuddy, 1983. *Evaluating the medical literature Part II: Statistical Analysis*, Ann. Emerh. Med., 12: 610-613.

Freeman, J.A., T.C. Chalmers, H. Smith, *et al.,* 1978. *The importance of Beta, the type II error and sample size in the design and interpretation of randomized clinical trial*, New Engl. Jr. of Med., 299:690-694.

Gupta, S.C., and Kapoor, V.K. 1970. *Fundamental of mathematical statistics*, SC Publications, New Delhi, India

Hopkins, K.D. and Glass, G.V. 1978. *Basic statistics for the behavioral Sciences*, Englewood Cliffs, New Jersy, Prentice, USA.

Keppel, G., 1978. *Design and analysis. A researcher's handbook*. Englewood Cliffs, New Jersy, Prentice, USA.

McCullagh, P.R and Nelder, J.A. (1989). Generalised linear models (second edition). pp. 511, London: Chapmanner Hell.

Rao, N.S.N. 1985. *Elements of Health Statistics*, First edition, R. Publications, Varanasi, India.