RELATIONSHIP BETWEEN SOME STATISTICAL METHODS

D. SCOTT

Plant Physiology Division (Substation), D.S.I.R., Lincoln

INTRODUCTION

In most ecological problems we are endeavouring to define various aspects of plant/animal/environment interactions. The object of taking measurements is to:

(1) find the magnitude of particular effects or interactions and;
(2) to determine their reliability.

Statistical or probabilistic methods are required whenever a decision has to be made whether a particular conclusion can justifiably be deduced from a particular set of data after the variability, chance effects or errors inherent in the data have been considered. Having accepted that, most ecologists may be bewildered by the variety of statistical methods available. The following is an attempt to rationalize the differences between some of the methods and thereby to assist in selecting the one most suitable for a particular problem.

Another point which has influenced the style of presentation is the belief that, in this present era of computers and package programmes, the biologist is no longer limited by his knowledge of all the computational details of particular methods. Instead — as it should have been always — the main emphasis should be to understand what each method does or does not do, so that the most appropriate one is used.

A consideration of the nature and distribution of the values of the variables shows that each statistical method has been developed for a particular type of problem and that although there is little overlap between them, they form a continuum of methods.

The texts referred to in preparation of this paper were: Mood (1950), Hotelling (1954), Tukey (1954), Anderson (1958), Ezekiel and Fox (1959), Scheffe (1959), Turner and Stevens (1959), Williams (1959), Theil and Goldberger (1961), Cooley and Lohnes (1962), Lawley and Maxwell (1962), Pearce (1965), and the two most frequently referred to Crow, Davis and Maxfield (1960) and Seal (1964).

VARIABLES, MODELS AND RANDOMNESS

Variables

In most problems one will be primarily interested in a particular variable (*subject*), for example, plant or animal growth, and how this is influenced by other variables (*factors*). The terms 'subject' and 'factors' are used to emphasize that statistics is dealing with the relationship between numbers, even though biologically there may be definite functional relationships between the variables which these numbers represent. Furthermore, the 'subject' in one problem may be the 'factor' in another, and in other problems it will be impossible to distinguish between these alternatives. There is a degree of synonomy between these terms and others used in statistics, as follows:

subject: dependent variable, y, effect, and output variable.
factors: independent variables, x, cause, input variables, and explanatory variables.

The variables which are used may be defined with various degrees of accuracy:

(a) poorly-defined or unmeasurable, e.g. common sense or ecological amplitude.
(b) qualitative, e.g. male/female or species A/species B/species C.
(c) quantitative (1) discrete variables, e.g. instar 1/instar 2; week 1/week 2/week 3.
(2) continuous variables, e.g. height 3.4 cm., 9.8°C., pH 7.3.

Depending on the number, quantitative discrete variables may be treated as either qualitative or as continuous quantitative variables. Continuous variables may be coded and treated as qualitative or discrete variables, e.g.: Treatment 1 = 40 lbs./acre, Treatment 2 = 80 lbs./acre.

Models

All statistical techniques are based on a model or mathematical representation of the problem. For example, a model of the relationship between wheat yield, temperature and rainfall might be:

$$y = B_0 + B_1x_1 + B_2x_2 + e$$

where y is wheat yield; x_1 and x_2 appropriate

measures of temperature and rainfall; B_0, B_1 and B_2 constants; and a random variable (e) which has a mean of zero and a normal distribution of values independent of the other variables.

For a particular situation this relationship might be estimated by

$$y = b_0 + b_1x_1 + b_2x_2$$

where y, x_1 and x_2 have the same meaning as above and where b_0, b_1 and b_2 are estimates of the corresponding coefficients fitted according to the requirement of minimum least squares deviation between actual and estimated values of wheat yield.

It is the comparison between the model and the corresponding values estimated from the sample data which forms the basis of any deductions or conclusions. The relationships and conditions assumed in the model and the degree to which these are satisfied by the sample data determine the applicability of a method to a particular problem.

Most statistical techniques assume that the factor variables have separate and additive effects on the subject variable (linear model). This is often not so and the initial data may have to be transformed in various ways, or additional variables introduced to account for interactions, etc. to satisfy this requirement.

The decision on whether a model is appropriate to a particular set of data has to be made on the basis of the biological or ecological understanding of the problem. The statistician can give only the consequences of accepting a particular model.

Most techniques make only limited use of information from sources other than the data in hand, for example, from previous experience or experiments. The qualitative aspects of this information can be used in selecting which model or method is appropriate, but all quantitative aspects of the solution are usually obtained from the particular set of data. Some techniques (such as path analysis with mixed estimation) allow the use of previous quantitative information as well.

Randomness

All tests of statistical significance, statements of probability, etc. are based on a comparison of the values obtained as compared with those expected, assuming that the actual sample was obtained randomly. Although the conditions required may vary with the method, there will always be some conditions of randomness which will have to be met if tests of significance are to be made.

The computational procedures of any of the techniques can be applied even if the randomness requirement is violated and the method may still give the best estimates of the quantities sought (as in curve fitting). But where the randomness requirements are violated it is no longer possible to make a statistical test, estimates of confidence intervals, etc. Many parameters are of use only in the statistical context (e.g. correlation coefficients).

CRITERIA FOR DIFFERENTIATING BETWEEN METHODS

The various methods may be classified by consideration of the nature and distributions of the subject and factor variables, as follows:

(a) Nature of subject variables;
  (1) Qualitative,
  (2) Continuous quantitative.
(b) Nature of factor variables;
  (1) Unmeasurable,
  (2) Qualitative,
  (3) Continuous quantitative,
  (4) Mixture of quantitative and qualitative.
(c) Distribution of subject variables (assumed or preferably known);
  (1) Values have been randomly selected from a normally distributed set,
  (2) Values are not necessarily normally distributed. They may have been approximately selected by the investigator. However, the deviation between the actual and expected values of the variables are usually assumed to be normally distributed.
(d) Distribution of factor variables (assumed or preferably known);
  (1) Factors varying independently of each other and the values of each normally distributed,
  (2) Factors considered to have separate effects on subject variable. Otherwise there are no other restrictions on the distribution of the factor variable values, i.e. they may be selected values,
  (3) Factors may be correlated with each other and no restriction on the distribution of values.

The simplified relationships between the various methods according to these criteria are given in Table 1 and are further described and qualified in the next section.

TABLE 1: *Relationship between statistical methods in terms of nature and distribution of subject and factor variables.*

METHOD	REQUIREMENTS OF VARIABLES: FACTORS: unmeasurable	qualitative	qualitative & quantitative	quantitative	independent & normal dist.	independent	can be correlated	SUBJECT: qualitative	quantitative	normal dist.	not necessarily normal dist.
Path analysis	–	–	–	*	–	–	*	–	*	–	*
Multiple regression, Type 1	–	–	–	*	–	*	–	–	*	–	*
Multiple regression, Type 2	–	–	–	*	*	–	–	–	*	*	–
r (Simple correlation)	–	–	–	*	*	–	–	–	*	*	–
Mult. R (multiple correlation)	–	–	–	*	*	–	–	–	*	*	–
Partial correlation	–	–	–	*	*	–	–	–	*	*	–
Discriminant analysis	–	–	–	*	*	–	–	*	–	*	–
Covariance analysis	–	–	*	–	*	–	–	*	–	*	–
Analysis of variance, Type 1	–	*	–	–	–	*	–	–	*	–	*
Analysis of variance, Type 2	–	*	–	–	*	–	–	–	*	*	–
Contingency table	–	*	–	–	–	*	–	*	–	–	*
Principle component analysis	*	–	–	–	–	*	–	–	*	–	*
Factor analysis	*	–	–	–	–	–	*	–	*	*	–
Canonical correlation	*	–	–	–	–	*	–	–	*	*	–

DESCRIPTION AND COMMENTS ON METHODS

The methods are discussed in order of their possible familiarity to readers rather than in the logical order given in the table.

Contingency table:

This method provides a test for determining whether the frequency of combinations of qualitative characteristics in two or more categories differ significantly from the combinations that would be expected if, in fact, the characters in the various categories varied independently: As, for example, in testing for association between species from their presence or absence in a random set of quadrats. No distinction is made between subject and factor variables.

Analysis of variance, Type 1:

In its simplest form the method is used when the subject variable is quantitative and the factor variables are qualitative. The method partitions the sums of squares of deviations and degrees of freedom into components associated with particular factors, thereby determining the mean effect of the factor variables on the subject variable: As, for example, in the relationship between quantitative growth measurements of a species and the combinations of nutrient or diet where these can be described only qualitatively and where the combinations used are considered to be the only ones of interest. Statistical tests are based on the assumption that residual error effects are random, independent and normally distributed.

Since no assumption is made about the distribution of the values of the factor variables, these may be chosen by the investigator. Quantitative factor variables are treated in a qualitative manner. It is Type 1 analysis of variance problems which is usually discussed in elementary texts. The analysis of variance approach is often used in presenting the results of other methods.

Analysis of variance, Type 2:

This is similar to Type 1 but with the alternative restriction that the levels of the factors are themselves random samples from a normal population of values, as would be so if, in the above example, the diets were considered to be a random sample of all possible nutrient combinations.

The objectives in this instance are usually the estimation of the relative variation contributed by each factor rather than the estimation of mean

effects. Many practical problems are a mixture of Type 1 and Type 2.

Multiple regression, Type 1:

This determines values and confidence intervals of coefficients in an equation fitted by least squares between quantitative variables. The assumption is that the departure between estimated and actual values of the subject variable are randomly, independently and normally distributed. An example is the relationship between quantitative measurements of plant growth and quantitative measurements of climatic factors, when the levels of the climatic factors are selected and vary independently of each other (as in controlled climate studies).

Because there is no requirement that the values of the factor variables be normally distributed, they may be functions of other variables (e.g. polynomials), thereby allowing fitting of curvilinear and interaction relationships. However, the fact that some of the variables may be correlated is not taken into account in the equation-fitting procedure. Stepwise procedures allow factors to be included or excluded from the regression equation in the order of their significance. By using code variables to designate particular treatments (e.g. 0, 1) the method may be extended to include qualitative variables.

Multiple regression Type 2:

This is similar to Type 1 in fitting linear equations between quantitative variables but there is an additional requirement that the values of each of the variables (both subject and factor) be independently and normally distributed in addition to the linear relationship between them. This approach might apply to the previous example if, instead of selecting the levels of the climatic factors, they were assumed to be random samples of possible levels and combination and it was further assumed that there was no interaction between them.

Simple, multiple and partial correlation coefficients are part of this method. Curvilinear relationships and other types of interaction between factor variables cannot be considered.

Coefficient r (Simple correlation), Multiple R. (Multiple correlation coefficient):

This is a standardized index which shows the degree of relationship between pairs (or groups) of variables. To attach significance to the indices so obtained one must assume that the variables are random samples from a normal distribution of values.

No distinction is made between subject and factor variables.

Partial correlation coefficient:

For a group of variables each having a normal distribution of values, this is a standardized index which shows how the values of two variables change relative to each other when the effects of correlation with other variables in the set are removed.

Partial correlation should be distinguished from simple correlation when dealing with two variables from a set of several variables. Partial correlation systematically eliminates, whereas simple correlation merely ignores, the variation related to the correlation with the other variables.

Path analysis:

Like the multiple regression method described above, path analysis fits a linear equation between quantitative variables on the assumption that random normal departure exists between estimated and actual values of subject variables. However, in this method the factor variables in one equation may be the subject variables in an associated equation. This interaction between variables is taken into account in the curve-fitting procedure. For example, in the previous illustration between plant growth and climate, path analysis would be appropriate if consideration of the interaction between sunshine and temperature was required as well as the interaction between them both and plant growth.

Because interaction between factors is allowed for, a large number of interacting factors may be considered simultaneously. The relationship between the variables has to be derived from consideration of the biological or ecological issues involved. In addition, the method can deal with curvilinear and feedback relationship between the variables. An extension of the method allows the use of *a priori* quantitative data, thereby offering one method of synthesizing information from different sources.

Covariance analysis:

This technique is similar to analysis of variance in partitioning sums of squares and degrees of freedom of subject variables into components associated with each factor variable. It differs in that some of the factor variables are qualitative and others quantitative; as for example, in the relationship between quantitative measurements of animal body weight, feeding rations (qual.), climate (quant.), sex (qual.) and initial weight (quant.).

The method is usually described in terms of removing the effects of uncontrollable but measurable variables in problems falling under the heading of analysis of variance, as, for example, in the relationship between body weight, feeding and climate after removing the effects of sex and initial weight. But more generally it can be thought of as a way of dealing with problems in which there are both quantitative and qualitative factor variables. Most texts assume that the factor variables are random samples from a normal distribution of values. However, presumably there is a similar differentiation into Type 1 and Type 2 models as in analysis of variance and multiple regression.

Discriminant analysis (*simple and multiple*):

Given the quantitative characteristics (factors) of individuals in two or more groups (qualitative subject variables), discriminant analysis determines the relative weight which should be ascribed to each of these characteristics when attempting to assign a further individual to one of the groups. It does this by forming a linear function of the factors which defines an index whose values and confidence intervals are indicative of particular groups. An example might be that of assigning a site to one (or more) associations (a qualitative characteristic) from quantitative measurements of environmental factors, using the results of previous measurements made on examples of those associations.

The method can be used to discriminate between more than two groups. It is also possible to get subsidiary indices from the same set of data for more precise assignment.

Principal component analysis:

Given a group of measurable variables this method determines a second set of variables — linear combinations of the first set — which would also account for the observed variation. Such cases occur when the observable variables (subjects) are assumed to be the result of a number of poorly-defined unmeasurable factor variables and where one wishes to estimate the minimum number of mutually independent variables which would be necessary to account for the observed variation. Component analysis isolates the same number of factor variables as observable subject variables, though only some will be significant. For example, given the relative abundance of a number of species in a series of samples, principal component analysis may be used to define variables which define gradients (presumably environmental) along which the species are distributed.

Since, by definition, the factor variables are unmeasurable the strategy adopted is to define the estimated factor variables so that the maximum effect is contributed by a single factor; so that the second factor is independent of the first and accounts for the maximum of the remaining variation, and so on. In practice one assumes that it will be possible to give a biological meaning to the factors so defined.

Factor analysis:

This is similar to principal component analysis but differs in defining a specified number of factor variables which is less than the number of measured subject variables. Thus, given quantitative measurements of population fluctuations of several species within an area, it may be used to define factors which may, for example, reflect common density-independent factors.

The factors defined may be correlated with each other and need not be linear. However, within these restrictions there is a variety of different procedures which may be used to define the factors. Both methods are useful when nothing else is known about the relationship between the observed variables.

Canonical correlation:

This is concerned with the determination of linear combination of each of two sets of variables such that the correlation between the two linear combinations is a maximum. The relevant situation may arise when two sets of measurements of different types are taken on the same objects (e.g. morphological dimensions and intelligence tests) and when both are believed to be the result of one (or more) underlying factors. It is similar to component analysis and factor analysis in defining the underlying variables in terms of the observable subject variables. For example, given physical and chemical properties of a site, canonical correlations could be used to define which combinations of the former correlate with which combinations of the latter, thereby indicating the possible structural and functional relationship between the two.

Again, one hopes that a biological interpretation can be given to the correlation function so defined. Under certain conditions, when there is only one

variable in one of the sets, this method may be equivalent to either multiple regression or discriminant analysis.

DISCUSSION

In all types of ecological research we are faced with making decisions about the relationship between two or more variables on the basis of a sample of numerical data. Table 1 and the previous section are an attempt to briefly describe the requirements of each method, what they do, and the differences between methods. Both may then be used as a guide in selecting the method appropriate to a particular project or set of data. This could be done by, *firstly*, deciding which are the subject and factor variables; *secondly*, by checking each variable against the categories in Table 1 (a/ nature of subject variables, b/ distribution of subject variables, c/ nature of factor variables and d/ distribution of factor variables) to determine which techniques might be appropriate to the particular data; *thirdly*, by comparing these with brief descriptions given in the previous section; and *finally*, by consulting detailed descriptions of the appropriate method.

The methods may be divided into four groups according to the nature of the factor variables. In the first group these factors are continuous quantitative variables; whereas in covariance analysis and some forms of Type 1 multiple regression the factors may be a mixture of both qualitative and quantitative variables. Analysis of variance and contingency tables deal with qualitative factor variables but in factor analysis and the two associated methods the factor variables cannot be defined except in terms of the observed subject variables. Within each of these groups the methods may be further subdivided according to the nature of the subject variable (quantitative except in discriminant analysis and contingency tables) and on whether or not the distribution of the subject and factor variables are separate and normally distributed. In some methods there is no distinction between subject and factor variables (simple, partial and multiple correlation).

The description of the methods has been given mainly from the point of view of the user, as determined by the type of information which can be put into a method (e.g. quantitative, qualitative, etc.), and the type of information required in return. There has been little discussion of the models or null hypotheses on which each technique is based (apart from the restriction they impose on the techniques which are applicable) and no detailed discussion of the mathematical theory or procedures of computation. This gives a somewhat one-sided view in that, mathematically, most of the techniques described are really parts of the same general technique (least squares linear model) in which there is no distinction between qualitative and quantitative variables. This is well demonstrated by Seal's (1964) book. The same book also emphasises point made earlier, that the results obtained are dependent on the biological concept of the issues involved in a particular problem, e.g. Seal gives five alternatives and reasonable sets of hypotheses for one set of data.

There is an inverse relationship between the order in which the methods are given in Table 1 and the order in which they might be applied to a particular problem. In the first stages factor analysis or associated methods may be used to indicate what relationships might exist between observable variables. Progressively, as the relevant factors are identified — first qualitatively and then quantitatively and their interactions with other factors delineated — the appropriate statistical technique will progress through analysis of variance, correlation, multiple regression and path analysis. Of the methods discussed path analysis indicates the most desirable state of affairs, since its use implies (i) that all variables have been measured quantitatively with no restriction on their values and (ii) that the probable relationships between the variables is known and that some quantitative information may be included about these relationships. Furthermore, I regard any method which requires a normal distribution of values of factor variables as being less desirable than an equivalent method without such a requirement (Type 2 versus Type 1 problems). This is because I believe that any small advantages that there may be in being able to determine other parameters by assuming normality is outweighed by the practical disadvantages they impose on the sampling procedure. For example, the determination of relationship between forest growth and environmental factors, using the normality assumption, would require that most of the sampling effort be expended in repetitious sampling near the modal conditions rather than in the deliberate sampling of a range of conditions.

Finally, to reiterate the main point: The biologist no longer needs to be conversant with all the

details of computation of any of these methods, but it behoves him to be aware of their essential features so that he chooses the one appropriate to the particular problem. I hope that this paper is a contribution to that end.

REFERENCES

ANDERSON, T. W. 1958. *An introduction to multivariate statistical analysis*. Wiley, New York.

COOLEY, W. W. and LOHNES, P. R. 1962. *Multivariate procedures for the behavioral sciences*. Wiley, New York.

CROW, E. L., DAVIS, F. A. and MAXFIELD, M. W. 1960. *Statistics manual*. Dover, New York.

EZEKIEL, M. and FOX, K. A. 1959. *Methods of correlation and regression analysis* (3rd edition). Wiley, New York.

HOTELLING, H. 1954. Multivariate analysis. *in* KEMPTHORNE, O., BANCROFT, T. A., GOWAN, J. W. and LUSH, J. L. (eds.) *Statistics and mathematics in biology*, pp. 67–80. Iowa State College Press, Ames.

LAWLEY, D. N. and MAXWELL, A. E. 1962. Factor analysis as a statistical method. *Statistician 12*: 209–229.

MOOD, A. McF. 1950. *Introduction to the theory of statistics*. McGraw Hill, New York.

PEARCE, S. C. 1965. *Biological statistics*. McGraw Hill, New York.

SEAL, H. L. 1964. *Multivariate statistical analysis for biologists*. Methuen, London.

SCHEFFE, A. 1959. *The analysis of variance*. Wiley, New York.

THEIL, H. and GOLDBERGER, A. S. 1961. On pure and mixed statistical estimation in economics. *Intern. Econ. Rev.* 2: 65–78.

TUKEY, J. W. 1954. Causation, regression and path analysis. *In* KEMPTHORNE, O., BANCROFT, T. A., GOWEN, J. W. and LUSH, J. L. (eds.) *Statistics and mathematics in biology*, pp. 35–66. Iowa State College Press, Ames.

TURNER, M. E. and STEVENS, C. D. 1959. The regression analysis of causal paths. *Biometrics* 15: 236–258.

WILLIAMS, E. J. 1959. *Regression analysis*. Wiley, New York.