ANIMAL- OPEN SPACE

STATISTICAL GUIDE FOR AUTHORS

DESIGN

Important points to consider when planning an experiment are: the amount of replication, the method of randomization, and the system of blocking (if any) to be used. Blocking allows major components of variability to be eliminated from estimates of treatment effects and experimental error.

Randomization ensures unbiased comparisons and estimates of error variance. Replication should be sufficient to ensure that treatment differences of practical importance can be detected as statistically significant.

Simple designs (completely randomized, randomized block, split-plot) cater for most experimental situations. Seek the help of a statistician if a more complicated design seems necessary.

The power of a test is useful at the planning stage in determining the required size of an experiment or sample. For example, if it is required that the power of a test is at least 0.9 when the true effect is some given value, it is possible to calculate the necessary size of experiment or sample size (Lynch and Walsh, 1998, Appendix 5). However, power calculations have no role in the analysis of data.

ANALYSIS OF VARIANCE

The main use of analysis of variance is the calculation of the residual variance for an experimental design. Variance ratio (F) tests for various hypotheses associated with the design (not necessarily the hypotheses of interest) are produced as a by-product. This is straightforward when data are balanced (when there is the same number of observations for each treatment or treatment combination). The analysis of unbalanced data requires more care. With two factors A and B, and unequal replication of the different combinations of levels of A and B, effects and sums of squares for B need to be adjusted for the effects of A, and vice-versa.

VARIANCE COMPONENTS

Analysis of variance can also be the basis for variance component estimation. Again, this is straightforward when data are balanced. With unbalanced data, the ANOVA table is not unique, and it is preferable to use another method, such as restricted maximum likelihood (Cox and Solomon, 2002).

REGRESSION

The dependent variable Y is random with expectation a linear function of X (the independent variable). The values of X are often chosen by or under the control of the experimenter, while the values of Y are subject to experimental error. Sometimes both X and Y are random, in which case there are two regression lines, Y on X and X on Y. The first is relevant if it is required to predict Y from X, and vice-versa. An exception to this general rule occurs in calibration problems, where Y is measured at a series of fixed values of X, and the fitted line is subsequently used to predict X from Y.

When data are grouped, it may be necessary to separate between-group and within-group regressions. These may be quite different.

Stepwise selection procedures may be used to reduce an unwieldy set of independent variables to a more manageable subset. There may be a large number of subsets, all giving approximately the same quality of fit. If the purpose of the regression is prediction, it may not matter which subset is chosen. If the purpose is scientific understanding, it is more important that the reduced equation be biologically sensible than that it be chosen by a statistically optimal procedure.

REPEATED MEASUREMENTS

Repeated measurements on the same experimental unit should not be regarded as independent. Sometimes a split-plot design with experimental units as main plots can be used. In growth studies, a simple method of analysis is to do a separate analysis on a series of summary measures for each experimental unit (e.g. mean, linear, quadratic trend).

TRANSFORMATIONS, GENERALIZED LINEAR MODELS

Special methods may be required if the data are skew or otherwise non-normally distributed, if the residual variance is not constant, or effects (treatment differences) are not constant on the original scale of measurement. Note that it is the residuals obtained after fitting regression or ANOVA which should be normally distributed, not the raw data. One approach is based on transformation of the data. Some commonly used transformations are the square root, for data in the form of counts, and the arcsine transformation, for binomial data. If the variance seems to vary as the square of the mean, a

logarithmic transformation may be useful. Standard errors are not constant for comparisons on the original scale, so estimates, standard errors, and test results are best given on the transformed scale. The equivalent effects on the original scale are most easily given in the form of confidence intervals, obtained by back transformation of the end-points. An alternative to analysis on a transformed scale is to fit a generalized linear model (Dobson, 2002). This allows the problems of variance heterogeneity and additivity of effects to be tackled separately. The two approaches usually produce similar results.

ESTIMATION AND TESTING

Do not confuse standard deviation and standard error. The standard deviation is a measure of variability in sample or population. The standard error is a measure of the precision of an estimate.

Multiple comparison procedures, such as Duncan's multiple range test, Scheffe's test, or Bonferroni adjustment of P-values, are relevant when a hypothesis is to be treated as one of a larger set, or when it has been selected in the light of the data. However, a well-designed experiment sets out to test a small number of clearly stated hypotheses. The treatments are usually structured , e.g., a combination of factors each with a small number of qualitative levels, or quantitative levels varying on a continuous scale. In this case comparisons are determined a priori and there is no need for multiple comparison tests.

Statistical significance should not be confused with practical, or biological, significance. A small real effect, of no practical importance, may be statistically significant in a very large sample. A nonsignificant result does not demonstrate that there is no effect. It means that the data are compatible with there being no effect, and in small samples, this can happen even when the real effect is large.

Computer-intensive methods, such as Gibbs sampling (Gilks et al, 1996, Chapters 1 and 2) and bootstrapping (Davison and Hinkley, 1997), can be useful in providing standard errors, confidence intervals, or significance tests in non-standard problems when conventional methods fail. However, these methods are easy to misuse. Use with caution.

PRESENTATION

The statistical report should omit extraneous detail, but be informative enough to allow the reader to make independent judgments wherever possible. For example, in a regression analysis, plotting the raw data together with the regression line allows the reader to assess the need for transformation or the existence of outliers. Giving standard errors and degrees of freedom with estimates allows readers to choose their own significance levels for hypothesis tests or confidence intervals.

Do not over-use hypothesis tests. If an estimate is several times greater than its standard error, or is small relative to the standard error, a test may be superfluous. An estimate with standard error, or confidence interval, is often more useful and gives more information.

There are many statistical software packages, each with its own strengths and weaknesses. Do not assume that everyone is familiar with your chosen statistical package. In particular, avoid terminology which is specific to a particular brand of software. For example, terms such as 'least-squares mean' and 'type III sum of squares' may not be generally understood.

Observations which seem to be inconsistent with the main body of data should not be excluded from the analysis without good reason. If in doubt, it may be useful to analyze the data both with and without an anomalous value to assess the sensitivity of the analysis to its presence. In any case, omission of outlying values should be reported. More generally, any shortcomings in design or analysis should be reported with an indication of the possible effect on the results.

Round values to a reasonable number of decimal places. For example, an estimate given as 5.3125 with standard error 1.7082 has too many decimal places, and might as well be given as 5.3 with standard error 1.71. The standard error is usually given with one more decimal than the estimate. It is rarely necessary to have more than 2 or 3 significant digits.

REFERENCES

Cox, D.R. and Solomon, P.J. (2002). Components of Variance. Chapman and Hall/CRC, Boca Raton, Florida.

Davison, A.C. and Hinkley, D.V. (1997). Bootstrap Methods and their Application. Cambridge University Press.

Dobson, A.J. (2002). An Introduction to Generalized Linear Models. Chapman and Hall, London.

Gilks, W.R., Richardson, S., and Speigelhalter, D.J. (1996). Markov Chain Monte Carlo in Practice. Chapman and Hall, London.

Lynch, M. and Walsh, B. (1998). Genetics and Analysis of Quantitative Traits. Sinauer Associates.