DOI: 10.1111/j.1365-3180.2009.00758.x

Current statistical issues in *Weed Research*

A ONOFRI*, E A CARBONELL†, H-P PIEPHO‡, A M MORTIMER§ &
R D COUSENS¶

Department of Agriculture and Environmental Sciences, University of Perugia, Perugia, Italy, †Biometrics and Informatics Unit, Instituto Valenciano de Investigaciones Agrarias, Moncada (Valencia), Spain, ‡Universität Hohenheim, Department of Crop Production and Grassland Research, Bioinformatics Unit, Stuttgart, Germany, §School of Biological Sciences, University of Liverpool, Liverpool, UK, and ¶School of Land and Environment, University of Melbourne, Parkville, Victoria, Australia

Received 28 November 2009
Revised version accepted 28 November 2009

Summary

The correct design of experimental studies, the selection of the appropriate statistical analysis of data and the efficient presentation of results are key to the good conduct and communication of science. The last Guidance for the use and presentation of statistics in *Weed Research* was published in 1988. Since then, there have been developments in both the scope of research covered by the journal and in the statistical techniques available. This paper addresses the changes in statistics and provides a reference work that will aid researchers in the design and analysis of their work. It will also provide guidance for editors and reviewers. The paper is organised into sections, which will aid the selection of relevant paragraphs, as we recognise that particular approaches require particular statistical analysis. It also uses examples, questions and checklists, so that non-specialists can work towards the correct approach. Statistics can be complex, so knowing when to seek specialist advice is important. The structure and layout of this contribution should help weed scientists, but it cannot provide a comprehensive guide to every technique. Therefore, we provide references to further reading. We would like to reinforce the idea that statistical methods are not a set of recipes whose mindless application is required by convention; each experiment or study may involve subtleties that these guidelines cannot cover. Nevertheless, we anticipate that this paper will help weed scientists in their initial designs for research, in the analysis of data and in the presentation of results for publication.

Keywords: statistics, data, design, analysis, ANOVA, regression, models, multivariate analysis, significance, probability, biometrics.

ONOFRI A, CARBONELL EA, PIEPHO H-P, MORTIMER AM & COUSENS RD (2010). Current statistical issues in *Weed Research*. *Weed Research* **50**, 5–24.

Introduction: philosophy and limitations of statistics

Statistical methods are used to assist scientists in interpreting data, which are, by the very nature of biological systems, variable. For this reason, the use of statistics has become an indispensable part of the process of research.

On the one hand, this is critical because the use of inappropriate methods can lead to misinterpretation of results and faulty conclusions. On the other hand, surveys of biological and agricultural journals in the past have shown that up to 70% of articles either use, or report, statistics incorrectly (Johnson & Berger, 1982). More recently, similar, although less dramatic, results have been reported for journals in the medical field (Clayton, 2007). It is clear that meeting high standards should be required in all research and publication efforts and statistical analyses should be regarded as a part of the research itself and treated with the same rigour as experimental methods.

Correspondence: Andrea Onofri, Department of Agriculture and Environmental Sciences, University of Perugia, Borgo XX Giugno, 74-06121 Perugia, Italy. Tel: (+39) 075 5856324; Fax: (+39) 075 5856344; E-mail: onofri@unipg.it

Within the scope of *Weed Research*, statistical analyses should help to communicate to the reader the meaning of experimental results and should neither distract from the biological explanation, nor make the text turgid and difficult to follow, or tables and figures less easy to interpret. Statistical accuracy does not necessarily imply complexity and difficulty. On the contrary: excessively complex analyses may indicate that the experiment was poorly conceived at the outset, the objectives were poorly defined or the researcher tried to fit too many things into one experiment.

The present contribution is meant to improve and standardise the level of statistical information reported throughout *Weed Research*. This document stems from the guidance previously published in this Journal (Anonymous, 1988) and tries to pursue a slightly broader aim. In detail, *Design and statistical input* section deals with the correct design of experiments and shows how the decisions taken at this stage should always be reflected in the selection of statistical methods. *Appropriate use of traditional techniques:* ANOVA *and regression* section takes into consideration the 'traditional' methods of ANOVA and regression, highlighting some issues that need to be appropriately accounted for, when using these widespread techniques. Even though ANOVA and regression have played an important role in weed science for a long time, several relatively new techniques may be more effectively used in some common experimental situations, which are listed in *When do we need to go beyond traditional* ANOVA *and regression* section. *Several response variables: multivariate statistics* section considers multivariate techniques, which may be very useful in the case of multivariate 'entities', such as weed flora. *Presenting the data* section is more practical and it is meant to guide authors to the presentation of statistics in *Weed Research*, through a list of statements that partly summarise the concepts already exposed in the previous sections.

We do not expect that all authors are willing to read this whole document, before submitting a paper to *Weed Research*. Therefore, wherever possible, we have attempted to make each section and subsection relatively independent from one another, so that authors and editors may easily select and read only the most relevant parts, according to their needs.

Design and statistical input

The experimental design dictates the inferences that can be reliably drawn from an experiment. A proper design is thus a requirement, not only for field and glasshouse experiments, but for all kinds of studies: it has been clearly shown that also experiments in highly controlled environmental conditions are subject to random variability (Measures *et al.*, 1973). Furthermore, the decisions taken at the design stage should be reflected in the selection of statistical methods.

Authors and editors should carefully consider the following main questions:

1. Are the experimental units clearly defined?
2. Is any form of pseudoreplication clearly distinguished from true replication?
3. Has randomisation been applied correctly?
4. Is the presence of controls appropriately accounted for in the analysis?
5. Are structured treatments and/or relationships between factors appropriately recognised?
6. Are blocking units appropriately accounted for during data analysis?
7. Are successive measurements (in space or time) taken on independent experimental units? If not, has a mixed model approach been used to account for serial correlation?
8. Is the experiment independently repeated in space or time?

If you are in doubt about some answers, this may be a warning message: some appropriate action needs to be taken during data analysis. The following parts of this section may serve as guidance to ensure that the experiment has been correctly designed and statistical methods to analyse the data clearly follow from that design.

Experimental units, replicates, pseudoreplicates and randomisation

The *experimental unit* is the smallest unit to which the process to allocate the treatment in randomised order is applied. For example, if a pot of five plants is sprayed with an herbicide, the experimental unit is the pot and not each of the five plants. Experimental units should be independently chosen, otherwise any casual event influencing one of them will also influence all the others, making the 'treatment effect' indistinguishable from 'background noise'.

Experiments may need *replicates*. In this case, it is important to recognise the difference between a true replicate and a pseudoreplicate. We can talk about true replicates when the randomisation process to allocate the treatment is applied to several independent experimental units. This must be clearly distinguished from pseudoreplication (sub-sampling), where several measurements are taken on a single sample and thus they are not independent, because they share the same 'sample'. Some typical examples would be: (i) spraying a pot with five plants (as above) and measuring separately the weight of each plant, (ii) treating one soil sample with one herbicide and making four measurements of concentration on four subsamples of the same soil, (iii) collecting one soil sample from a field plot and repeating four times the same chemical analysis. In all the above cases, the treatments are applied only to one unit (pot or soil sample) and there are no true replicates, no matter how often the unit is sub-sampled.

Pseudoreplication should never be mistaken for true replication, even in the case of laboratory experiments (Morrison & Morris, 2000). Pseudoreplication is not in itself a mistake, but authors and editors should always make sure that the lack of true replicates is appropriate to the experiment (see e.g Plant, 2007). In all cases, psedoreplication requires the adoption of appropriate methods of analysis. The main problem is that sub-sampled units will not be independent from one another, which violates one of the basic assumptions for the use of traditional statistical methods (ANOVA or regression). Readers are referred to *Pseudoreplication and other grouped data* section for suggestions on how to deal with this issue.

Another basic aspect of experimental design relates to *randomisation*, which justifies the use of a model with independent errors and avoids biased estimates of effect sizes. Authors should make sure that randomisation is performed correctly in any kind of experiments, including controlled environment studies. However, many classical designs (randomised complete blocks, split-plot, etc.) constrain the complete randomisation of the experiment, which is not a problem, as long as constraints are taken into account in the analysis, including the proper terms (blocks, rows, columns, etc.) in the ANOVA or other models used for analysis.

Relationships between factors

All sources of variation that have been included in the design and affect the measured variable of interest are called factors. With this term, in this paper, we will refer both to quantitative and qualitative explanatory variables. To study their effect, factors should be allowed to vary. Each value is a level of the factor. Generally, treatments take the role of the levels of factors, although sometimes they refer to the factors themselves.

If the experiment has only one factor, its levels may be unrelated or they may have some internal structure (e.g. a set of increasing doses of herbicides); if reasons were strong enough to introduce such a structure, it is necessary to avoid (or at least justify) the use of any statistical methods that would ignore it (such as multiple comparison testing).

If the experiment has more than one factor, the relationships between them should be clearly stated in *Materials and methods* and data should be analysed accordingly. Typically, in a crossed factorial design, all the combinations between levels of factors are included in the experiment. In this layout, factors may interact and such an interaction should be taken into account in data analysis and interpretation of results. By contrast, a factor B is nested within another factor A (called main factor) if each level of B is represented in only one level of the main factor. If reasons were strong enough to introduce either a factorial or a nested design, it would not be appropriate to disregard these relationships during data analysis.

Type of effects

Factors are classified into fixed or random. The levels of a fixed factor represent either (i) all possible levels for that factor or (ii) the only levels of specific interest, about which inference is to be made. The levels of a random factor can be seen as randomly selected from a wider population of possible levels. By convention, an effect is taken as random if any of the factors involved is random. The process of data analysis should always account for the random or fixed nature of the different factors and, if the experimental design includes both types of effects, a mixed model approach should be followed. Some examples of factors that are often regarded as random are: environments (years, locations, glasshouses and growth chambers), blocks, plots, Petri dishes and subjects (experimental units) in general. Statistical textbooks can be consulted to check if effects are fixed or random (e.g. Maxwell and Delaney, 1990).

Controls

The decision to include a control in the experimental design should be justified through the objectives of the experiment, in the same way as any other treatment. In this case, authors may wish to formally compare the control with all the other treatments and thus they will include the control also in the process of data analysis. This is a correct practice, because it provides more degrees of freedom for the estimation of error variance; in the case of factorial designs, some extra modelling may be necessary to separate controls from the other treatments (Piepho *et al.*, 2006).

When including a control in the analysis, it is very important to make sure that variances are homogeneous. Indeed, the control may often show a very high (or very low) variance with respect to all the other treatments, which may lead to biased results, lower efficiency or the unnecessary adoption of a stabilising transformation (Ahrens *et al.*, 1990; Phelps, 1991). In this case, the control should preferably be erased from data analysis and, if the ranges of data clearly do not overlap, it may be acceptable to conclude that control and treatments differ, without a formal test of significance. A more advanced solution may be to fit a mixed model with heterogeneous variances between control and treatments.

Blocking units

Blocking techniques are often used to control the contribution of nuisance factors to error variability. Several forms of blocking have been available for a long time (complete blocks, Latin square, incomplete blocks, rows and columns; see e.g Cochran & Cox, 1957; John & Williams, 1995). If some of those forms of blocking are introduced into the experiment, this should be clearly mentioned and justified. In the case of a split-plot design, authors should explain the experimental layout for all error strata (main and subplots, for example), as different forms of blocking may be introduced in each stratum. For example, subplots can be completely randomised or laid out according to an incomplete block design. Similarly, main plots may be completely randomised, or laid out in complete blocks or in rows and columns.

Repeated measures and repeated experiments

In some cases, the same experimental unit is repeatedly measured with respect to a factor of interest (generally time or space). Some examples are: (i) weekly measurements of height to estimate growth curves, (ii) sequential harvests of perennial crops, (iii) daily recording the number of germinated seeds on a Petri dish, (iv) collecting samples at different depths on the same plot. These examples lead to the concept of repeated measures or longitudinal data (when measurements are taken over time). The concept is similar to subsampling, with the major difference that the repeated factor (time or space) is not randomly selected within the experimental unit, but it is ordered along a temporal or spatial metric. Therefore, observational units are not independent, but they may exhibit some autocorrelation pattern, that is also known as 'serial correlation'. Some advice to deal with serial correlation is given in *Repeated measures/longitudinal data* section.

Similar to the concept of repeated measures, we can mention the concept of repeated experiments, when the whole experiment is repeated in a different time or place. This does not pose relevant problems in terms of data analysis and it should be considered a mandatory practice, as implied by the guidelines for authors of *Weed Research* (manuscript types): 'Research should cover sufficient temporal and spatial variation to be able to make sound generalisations'. Such a statement does not only refer to field and glasshouse trials, but also to experiments carried out in more controlled conditions, with material that is subject to spatial and temporal variability (seed populations, batches of treated soils, etc.).

Appropriate use of traditional techniques: ANOVA and regression

The most common experimental situation in weed science is represented by one or more fixed treatment factors and one quantitative response variable. In this situation, the usual action would be to fit some sort of linear or non-linear model, i.e. a regression model (when the treatment factor is quantitative), or, in the common case of experiments with replicates, to perform an ANOVA.

These techniques are very well established among weed scientists, but in our experience they are not always used appropriately. It is necessary that authors (and thus editors) make sure that all the following issues are appropriately considered:

1. Did you check for the basic assumptions after performing an ANOVA and/or regression?
2. If necessary, did you take the appropriate correcting measures?
3. If you performed contrast or multiple comparison testing, does this clearly fit within your experimental design?
4. If you perform linear/non-linear regression, did you check the final model for lack-of-fit?
5. If you compared regression curves or built complex regression models, were your decisions based on the appropriate statistics?

The following part of this section is aimed at guiding you towards an appropriate use of traditional statistical techniques.

Checking for basic assumptions. Outliers

It should never be forgotten that ANOVA and regression (as well as any other type of linear model) make assumptions and it is necessary to ensure, as far as possible, that these assumptions are satisfied. Without this basic check, it is not possible to guarantee that results are reliable and unbiased.

The three main distributional assumptions that should always be verified relate to normality, homogeneity of variances (homoscedasticity) and independence of experimental errors. Possible outliers should also be inspected, as they may adversely affect parameter estimation and inference.

The lack of independence in weed science may arise when observations are grouped, such as in case of pseudoreplication, repeated measures, split-plot designs and so on. Such grouping should be appropriately accounted for in the analysis, as shown in *Pseudoreplication and other grouped data* section.

The check for the other basic assumptions is generally performed 'ex-post', i.e. the selected model (e.g. ANOVA or regression) is fitted and then inspected. The lack of normality and homoscedasticity, as well as the presence of outliers, affect the distribution of residuals and thus a graphical inspection of these latter may be crucial. In the common case of fixed effects models, a graph of 'residuals vs. predicted' and a quantile–quantile (Q–Q) plot may suffice, even though more advanced methods exist, that are thoroughly discussed, for example, in Faraway (2004). In any case, authors should always state whether basic assumptions were carefully checked and how. This is particularly important (i) with counts and proportions based on a small number of replicates, which can not be assumed as normally distributed and (ii) when results differ by more than an order of magnitude, so that their variances may not be homogeneous.

Correcting measures: transformations

Once outliers have been carefully inspected and appropriate action (removal or correction) has been taken, large remaining deviations from normality and homoscedasticity require a correcting measure. If no measure is taken, reasons should be given in the Materials and methods.

The simplest action is to adopt a suitable transformation of the response variable, chosen by theoretical considerations or previous experience. Instead of making an arbitrary selection, authors may consider several families of transformations, such as the Box and Cox (1964) family. Other types of transformations are reviewed in Atkinson (1985) and Piepho (2003).

Even though stabilising transformations represent a useful and mathematically simple solution to non-normality and variance heterogeneity, they may result in several complications during the interpretation and presentation of results, mainly because transformed variables lose their direct biological meaning. As the main consequence, after performing an analysis on transformed data, authors are faced with the problem of selecting among three alternatives: (i) report the means of transformed data; (ii) report the naively back-transformed means; (iii) report the means on the original scales, even though analyses were performed on transformed data. Each choice has advantages and drawbacks that are discussed in *Presenting transformed data* section.

Because of the above complications, some care should be taken with reference to the following issues:

1. Do not routinely transform certain types of data: transformations may not be needed if no attempt is made to use a parametric test. For example, a graph of plant biomass over time may show different standard errors for means at each sampling date; this is not a problem as long as no attempt is made to ask whether plant sizes differ between dates. Additionally, consider that linear models (ANOVA, regression, etc.) are quite robust to moderate departures from normality and homoscedasticity;
2. If possible, try to be consistent throughout the paper in the choice of transformation;
3. Make sure that the transformation was actually effective to meet the basic assumptions for linear models;
4. In case of regression models (especially non-linear models), do not forget that transforming the response will distort the original relationship between the response and the predictor, so that model parameters will no longer retain their original biological meaning. To avoid this, consider a transform-both-sides approach, where both the observed data (left-side of model) and predictor (right-side) are transformed (Carroll & Ruppert, 1988). Several examples of this approach may be found in the weed science literature (see e.g. Streibig, 1988);
5. If necessary, consider other correcting actions or other statistical methods more specifically designed to analyse certain types of data. For example, the heteroscedasticity may be corrected by using the method of weighted least squares, with weights that are inversely proportional to the variance at each factor level (Carroll & Ruppert, 1988). The possible dependence of the variance on the mean may be explicitly modelled by using a 'power-of-the-mean' or other variance models (see some examples in Ritz & Streibig, 2008). Heteroscedastic and correlated errors may be analysed by the mixed model approach, while non-normality and non-linearity, together with heteroscedasticity, may be appropriately treated by using generalised linear models (GLIMs). Another technique that has recently come into fashion is the use of robust estimators of the covariance matrix (sandwich estimators; Lumley & Heagerty, 1999).

Contrasts and multiple comparison testing

In case of experiments with replicates, the first step of data analysis will probably be an ANOVA, with the aims of (i) checking whether the above assumptions are met and (ii) estimating the pooled standard error ('pure' error). After the ANOVA, the next steps should diverge, according to the type of explanatory variables (factors) and to the nature of effects. A possible scheme might be as follows:

1. In the case of qualitative explanatory variables with fixed effects, authors may be specifically interested in the different treatment levels and thus may seek to compare suitably defined means (marginal means, cell means, etc.).
2. In the case of quantitative explanatory variables, the authors will probably be interested in the overall pattern of response and thus they will switch to linear/non-linear regression.

The most common way to make all pairwise comparisons of means (fixed effect factors) is through the adoption of Multiple Comparison Procedures (MCP). The approach is very handy, but it may create a 'multiplicity problem', because the number of statistical tests increases progressively, as the number of means to be compared increases.

There has been much debate in the past about the use and misuse of MCP; during the 1970s, MCP (and particularly the Duncan's Multiple Range Test; DMRT) were adopted as standard methods of analysis for all datasets. In response to criticisms, the use of DMRT was exchanged for the Student–Newman–Keuls and other more conservative tests. The debate went on (see also Cousens, 1988, in this journal) and nowadays, we have come to the point where some journals would not even consider papers where MCP are used.

We feel that the use of MCP should not in principle be considered right or wrong, it depends on the objective of the experiment:

1. It is fully justified in the case of experiments aimed at comparing a set of unrelated factor levels, for example a number of different brands of herbicides or crop varieties.
2. It is inefficient (but some would say it is wrong) in the case of a quantitative explanatory variable (e.g. a series of doses or times), where regression analysis and curve fitting procedures may be more appropriate (see later). However, there might be cases where the researcher may not be able to identify or successfully fit a regression model; in these cases, contrasting the means of the successive levels (doses or times) might be an appropriate alternative to curve fitting, as the last resort.
3. It is inelegant (illogical) and inefficient when MCP are used in experiments where treatments show some internal structure. For example, we may have one untreated control, one group of chemical herbicides, one group of physical methods of weed control and, within this latter group, we may have two solarization and two flaming methods. In this case, the questions inherent in the objectives of the research should be stated in terms of a few 'contrasts of interest' (Pearce, 1992), that directly stem from the internal structure of the experiment (in this case: untreated vs. treated, chemical vs. physical, solarization vs. flaming).
4. It is inelegant, when it is used in a poorly designed experiment or a poorly interpreted dataset, wherein hypotheses were not established in advance. A wider adoption of contrast methodologies, following a careful planning of experiments (as in the above example), might help to limit the use of MCP to the cases wherein it is fully justified.

In the end, we do not suggest avoiding the use of MCP at all, but we do encourage authors not to go straight ahead with them and to think carefully if they really need all pairwise comparisons of treatment means. If so, authors should also take a careful decision on which procedure they should adopt. An acceptable work plan may be:

1. Consider if you want to control the comparison-wise (CWE) or the family wise error rate (FWE). The CWE rate is the probability of wrongly rejecting the null hypothesis (type I error) in one individual test, while the FWE rate is the probability of at least one wrong rejection of the null hypothesis when performing multiple tests. A CWE-controlling test might be preferred when only a few comparisons or contrasts are to be tested, each having a strong biological relevance (single-contrast problems), while FWE should be preferred in the case of a relatively high number of simultaneous tests, especially when an overall conclusion tends to be wrong when at least one single test is wrong (Abdi, 2007). Do not forget that, for a given sample size, a higher protection against type I error implies a higher risk of type II error (wrongly accepting the null hypothesis).
2. If you just want to control CWE in pairwise comparisons, use *t*-tests [Least Significant Difference (LSD) for balanced data]. There is no simpler and more efficient alternative.
3. To control FWE, use the Tukey test that provides a critical difference (Honest Significant Difference; HSD) for balanced data. For unbalanced data, there might be better alternatives, such as the simulation-based procedure of Edward and Berry (1987), which practically coincides with the Tukey test in the balanced case, or procedures based on the general multivariate t-distribution (Hothorn *et al.*, 2008).
4. To control FWE, do not use Student–Newman–Keuls or Duncan's MRT, because they do not control the FWE and do not yield a single critical difference for balanced data (Hsu, 1996).

5. To compare a single level (generally a control or predefined treatment) to each and every one of the other levels of the factor, the Dunnett test should be used.
6. In a factorial experiment, all of the above will also apply to comparisons controlling for the level of one or more treatment factors. For example, in a two-factor experiment with factors A and B and presence of interaction, a comparison of A-means by levels of B, and vice versa may be of interest.
7. Do not forget that the same kind of 'multiplicity' problem is raised whenever multiple tests are performed within an experiment, thus also when testing several contrasts. In all these cases, an adjustment procedure should be used, such as the Bonferroni's method. Other and more advanced methods may be found for example in Hothorn *et al.* (2008).

Further information on MCP may be found in Hsu (1996) and there is a useful chapter by Maxwell and Delaney (1990). Referring to the use of critical differences and significance letters in the presentation of results, some advice will be given in *Presenting the data* section.

Linear and non-linear regression. Checking for lack-of-fit

Quantitative explanatory variables, for example a set of increasing herbicide doses, call for regression analysis. Indeed, even though we include in the experiment some particular dosage levels, we are actually interested in the shape and scale of the overall response to the increasing/decreasing dose (see e.g. Schabenberger *et al.*, 1999).

During the 1980s, polynomial regression used to be very frequent, whenever curves deviated from linearity (which is almost always the case in weed science). Nowadays, the availability of fast computers has triggered a very widespread use of non-linear regression techniques and the choice of the appropriate model (linear, non-linear, etc.) has become mainly a matter of biological validity, according to the aims and experimental design. In this respect, it is necessary to point out that polynomial regression may still play its role, thanks to its simplicity and to the possibility of providing a good description of the dataset (high R^2 values). However, care needs to be taken, in that the shape of the fitted curve may not be supported by the data and may not be biologically reasonable.

Whatever model is chosen, least squares analysis makes the usual assumptions of normality, homoscedasticity and independence that have to be carefully verified with the appropriate diagnostic tools. Apart from this, it is also necessary to verify whether the equation shows a good fit with the experimental data. This may be easily done by inspecting the residuals and, in case of trials with replicates, by using an F-test for 'lack-of-fit'.

It is necessary here to stress that the coefficient of determination (R^2 statistic) should not be systematically used as a measure of goodness of fit, especially in non-linear regression. Indeed, the R^2 value may be seen as the comparison of the residual sum of squares (SS) for a fitted model with the residual SS for a model with only the intercept. This latter may not be among the parameters being fitted; thus the R^2 value that is reported by regression packages may not be meaningful.

Secondly, the R^2 value depends on the range of observed response: the lower the range, the worse the R^2 will look, even when the model is appropriate. In contrast, if the number of parameters is too high compared with the number of observations, a bad model could also result in a high R^2.

Last, but not least, it is possible to obtain high R^2 values also when estimated parameters make no sense or confidence intervals are very wide. This may frequently happen for example in herbicide bioassays, when the observed range of response is too narrow to obtain good estimates of lower and higher asymptotes for a sigmoidal response model.

A traditional reference for regression analysis is Draper and Smith (1998). In the case of non-linear regression, a practical reference is Ritz and Streibig (2008).

An important general point regarding non-linear regression is that, as with linear and polynomial regression, the experimental design should be reflected in the overall model. For example, the effects of blocks or nesting factors are frequently disregarded, mainly because they are not easily introduced in non-linear least squares packages. In this case, non-linear mixed models may be necessary (see e.g. *Accounting for 'environmental' variability in repeated experiments* section) and an expert advice from a statistician may be very helpful.

Comparing regression curves and building models

In some cases, experiments are designed with increasing rates of a quantitative variable, measuring the responses at two (or more) levels of another qualitative explanatory variable. A trivial example might be the weight increase over time for several weed species: in this case time is the quantitative predictor, while the species is the categorical predictor.

In such conditions, it would not be efficient to compare the weight of those species at each observation time, by using a multiple comparison test. The correct approach here would be to compare the two regression curves or some of their parameters.

In the case of linear models, this may be done by using an ANCOVA approach, wherein qualitative predictors are incorporated within the regression model. A similar approach can be taken in the case of non-linear regression curves, where nested models may be compared by using a likelihood ratio test, as shown in the case of herbicide bioassays by Streibig *et al.* (1993).

Such an approach is very powerful and the process of model building may bring inside the fitted equation several explanatory variables, generalising the traditional approach of multiple regression. This raises the problem of variable selection and comparison of alternative models, that should not be done exclusively on the basis of statistical significance and/or R^2 values, to avoid the unreliable selection of 'the best model', biased parameter estimation and increased experimentwise error rate, due to multiple hypothesis testing. Otherwise, the information theoretic approach, based on Akaike's Information Criterion (AIC) or Bayesian Information Criterion (BIC), should be preferred, which introduces a penalty for the number of parameters in the model (Whittingham *et al.*, 2006).

When do we need to go beyond traditional ANOVA and regression?

Traditional ANOVA and linear/non-linear regression have been the most widely used techniques in agricultural research and, if appropriately applied, they may help solve a high proportion of statistical problems with great simplicity. However, their usage, if not technically wrong, may be inefficient in some important situations, such as:

1. Experiments repeated in several environments;
2. Pseudoreplication/subsampling and other types of grouped observations;
3. Repeated measures and longitudinal data;
4. Unbalanced designs;
5. Counts and proportions;
6. Rating scale data.

In these situations, it may be possible to perform correct analysis with traditional methods, but this may hinder a clear and effective presentation of results. As consultants, we do not routinely recommend to abandon ANOVA and regression, but we would like to highlight some possible alternatives, with particular reference to mixed models and GLIMs, that are becoming increasingly important in several branches of plant protection (Garrett *et al.*, 2004).

Mixed models contain at least one random and one fixed effect, plus the residual error. The concept is not new to weed scientists; for example, split plot designs are mixed models, because of the presence of two error strata (main plot error and residual error). Traditionally, mixed models have been analysed within the frame of ANOVA, but recent advances in both theory and computation have transformed the mixed model framework into a very flexible tool, for example to extend the power of linear models to grouped data with heteroscedastic and autocorrelated errors, both for balanced and unbalanced designs.

On the other hand, GLIMs may be seen as an extension of linear models to non-normal responses and they may be successfully used to analyse counts and other non-normal data, without a prior transformation. Mixed models and GLIMs can be combined as in generalised linear mixed models (GLMMs).

Compared with ANOVA and regression, mixed models and GLIMs will require some effort to be appropriately mastered; we will not go into any detail here, but encourage the interested readers to consult one of the available textbooks (e.g. Schabenberger & Pierce, 2002) or search for an expert's advice. The gain in reliability of results and clarity of presentation may be worth the effort.

Accounting for 'environmental' variability in repeated experiments

Experiments are frequently repeated in several 'environments' (locations, years, fields, batches, ovens, etc.), with the aim of measuring the amount of variability that they produce on the observed responses. So far, these repeated experiments have been frequently analysed by traditional ANOVA/regression, even though this often prevents an effective summarisation of results, especially in the case of non-linear regression.

As an example, we may think of the degradation curve of one herbicide in five locations; with non-linear regression, we can estimate five half-life values, one for each location and we can test for possible significant differences among them. In this case, the environment is regarded as a fixed effect and results specifically refer to the five levels under investigation.

On the other hand, it may be possible (and convenient) to regard the five environments as a random sample drawn from a population of interest; in this case, the resulting model has one fixed (time) and one random (environment) factor and, thus, it may be regarded as a mixed model. Within this framework, it is possible to estimate one average half-life value, plus one 'variance component', measuring the inter-environments variability, i.e. how the half-life changes from one environment to the other. This second model is more parsimonious than the former and allows for a more effective presentation of results.

Such an approach has been used for example in Nielsen *et al.* (2004) and it may be applicable whenever the design includes one or more random factors to account for 'environmental' variability, including blocks, sites and other blocking factors.

Pseudoreplication and other grouped data

If we compared five herbicides in a randomised block design with four replicates and measured the weight of surviving weeds on three randomly taken quadrats of 0.25 m^2 per each plot, we would have a typical 'subsampling' design. In this situation, it would be wrong to analyse the data as a randomised block with 12 replicates, because the data from the same plot are not independent from one another.

One possibility is to take the average of subsamples and submit this to further analysis. This is correct only if the number of subsamples is constant.

One further possibility is to analyse the whole dataset, making a clear distinction between the two error strata, i.e. true replicates (experimental units: plots) and subsamples (observational units: quadrats). This can be simply accomplished, also with low-level computing tools, by performing a split-plot ANOVA, with herbicides as 'main plots' and quadrats as 'subplots'. In this way, the treatment effect (main plots) is tested on the appropriate error term and observations taken on each plot are naturally grouped, which accounts for their autocorrelation.

A more advanced and flexible solution, that works equally well when the number of pseudoreplicates is not constant, is to switch to a mixed model statistical package and include the 'plot' as a random effect in the model (see Piepho, 1997, for an example).

All the above considerations hold in other cases where observational units are grouped within experimental units, i.e. plots, pots, Petri dishes, etc.; several examples are given in Pinheiro and Bates (2000).

Repeated measures/longitudinal data

As an example of longitudinal data, we may consider the analysis of growth for a crop at increasing weed densities, where successive measurements are taken on sub-samples of the same plot. The situation is similar to that described in the previous section about subsampling, with the important difference that sampling times are not randomly taken within plots, but they are ordered along a temporal metric. Therefore, we might expect that the autocorrelation pattern is partly explained by the relative position of sub-samples (serial correlation): for example, the closer in time are the sub-samples, the higher may be their autocorrelation. In this situation, it would be wrong to use a simple two-factor (weed densities and times) ANOVA with independent errors, because longitudinal data violate the independence assumption.

One possible solution, that is feasible also with basic level statistical software, is to regard the design as a split-plot, with densities on main plots and times on sub-plots and use a traditional ANOVA (split-plot in time). This solution groups the observations obtained in each plot (thus removing part of their autocorrelation), but it is overly simplistic, because: (i) it assumes equal correlation and it does not account for a possible decay of correlation over time (serial correlation); (ii) if we intend to use non-linear regression and fit a growth function to the observed data, the split-plot structure, with two error strata ('between plots' and 'within plots'), is not easily introduced, without using advanced statistical software facilities.

The most correct solution is to switch to a mixed model approach; in this case, we introduce the 'subject' (plot) as a random effect in the analysis and fit a growth function to the data obtained in each plot. Therefore, we obtain the estimated regression parameters, plus the estimated residual error and one 'variance component' for each regression parameter, measuring how the estimate changes from plot to plot. The advantage is twofold: (i) the analysis clearly reflects the experimental design, with all the necessary error strata, (ii) if necessary, several alternative structures (compound symmetry, general, autoregressive, etc.) may be introduced to account for serial correlation (Piepho *et al.*, 2004).

A similar situation occurs in weed surveys or spatial data, when measurements are repeated in different spaces (Dormann *et al.*, 2007).

Unbalanced designs

Apart from the case of one-factor completely randomised unbalanced designs, in all the other situations the unbalance requires particular attention with reference to the estimation of means and hypothesis testing. Unbalance arises when the number of replications for each treatment (or the number of combinations for each factor level in factorial designs) is not the same, because of the peculiarities of the experimental design or to data lost for unforeseen circumstances. In this latter case, the causes for the loss must be unrelated to the treatment, otherwise bias will be introduced.

A small unbalance has been traditionally corrected by manual estimation of the missing data (LeClerg *et al.*, 1962). Such a procedure may now be considered out of date, as fitting a linear model implicitly accounts for any unbalance, providing that all the relevant model terms are included.

Concerning the estimation of treatment means, in the case of unbalanced designs, arithmetic (or marginal) means across observations are biased and thus should not be used for comparisons. Dealing with fixed effects, adjusted (least squares) means are indeed preferable (a simple example of calculation is given in Shaw & Mitchell-Olds, 1993), while for random effects the outcome for a certain factor level may be predicted by using the best linear unbiased predictors (see Piepho, 1994 for an example).

Another element of concern relates to the fact that with unbalanced data there are no unique decompositions of SS and the sequential (type I) SS that are output by several 'traditional' ANOVA packages, become dependent on the order with which effects enter the model. For example, if we have a randomised block design with one missing observation, the SS for treatments (and related *F*-test) will depend on whether the treatment effect has been entered in the model before or after the block effect. It is therefore fundamental to select a proper sequence of effects, as explained for example in Searle (1987). In the previous case, the correct test for treatments is obtained by fitting blocks first, then adjusting treatment effect for blocks. Provided that effects are specified in the correct order and marginality restrictions are observed (such as: fit an interaction only after main effects have been fitted; fit a quadratic term only when the linear term has been added already), all hypothesis tests in linear models can be done using type I SS (Nelder, 1994).

Other authors suggest the use of type III SS in the case of unbalanced data (see e.g Milliken & Johnson, 1984). In this case, some care needs to be taken, for example with a factorial design with unequal subclass numbers. Here, type III SS provide tests for main effects in the presence of interaction that test hypotheses not depending on subclass numbers. Nelder (1994) pointed out, however, that, in the presence of interaction, a test of main effects is irrelevant and that one should test for main effects only when interaction is deemed absent. In these circumstances, it is necessary to reduce the model by dropping the interaction and either (i) use the type I SS with two sequences, i.e. fitting B first (and A second) would provide a test for main effect A adjusted for B, while fitting A first (and B second) would provide a test for main effect B adjusted for A, or (ii), which is equivalent in this particular case, use the type III SS. These tests are the most powerful and hypotheses would be independent of subclass numbers.

In conclusion, in case of unbalanced data the advice for authors is twofold:

1. Do not to uncritically rely on treatment means and ANOVA tables, as obtained from statistical packages. They may not be easily interpreted.
2. In presence of random effects, adopt a mixed model approach, wherein balanced and unbalanced data are considered within the same framework.

Mixed models: things to care about

Even though we have given just a few examples of their possible applications in weed science, it should be clear that the mixed model framework is very powerful. However, this is sometimes a disadvantage, because we may be encouraged to adopt unnecessarily complex models and variance structures. For those who are already acquainted with mixed models, we would like to recommend a possible working strategy, to avoid 'overfitting':

1. Make sure that the inclusion of random effects is really justified and make sure that the observed levels of an effect may reasonably be considered as a random sample drawn from the population of interest. For example, if we have organised an experiment in two very close locations, with the aim of obtaining site-specific information, we should not regard the 'location' effect as random.
2. Start with a simple mixed model, without the inclusion of particular variance structures and check the basic assumptions. In particular, graphs of the 'within-group' residuals, together with 'Q–Q plots' of the estimated random effects may give a reasonable indication about the presence of heteroscedasticity and autocorrelation of errors.
3. If a careful inspection of residuals suggests so, select an appropriate variance structure. One simple way of selecting between two alternative variance structures is to fit the model with both structures and compare the fit on the basis of model selection criteria like the AIC or BIC: the lower the value, the better the model.

In general, mixed models may be a complex subject; apart from the above-mentioned references, another good example-based book is Littell *et al.* (2006). Use of mixed models is one of the cases where expert advice might be justified.

How to deal with binomial counts

We may consider an experiment where we tested some herbicides in a randomised complete block design with three replicates and assessed the number of dead weeds (*d*) and total number of weeds (*n*) in each plot. Working with counts, we should not in principle assume normality and homoscedasticity and thus it is wrong to use an ANOVA or regression, without considering an appropriate correcting transformation.

In this example, counts have a natural denominator (n) and the possible outcomes are only two (dead/alive), so that we normally talk about 'binomial' counts. As the first and traditional option, these counts are converted into proportions (d/n) or percentages and submitted to standard analyses, possibly after a suitable transformation of the response variable (square root or angular).

Such an approach is generally acceptable, but it may not be always efficient. First of all, counts may be positively skewed, with several 0s, which prevent an effective transformation into normally shaped data. Secondly, transformations will alter the scale of the dataset, making it less clear and readable (the arcsine square root of the proportion of dead weeds is not a natural metric for interpretation!).

Instead of trying to force the dataset to meet the basic assumptions for linear models, we might reasonably assume that it follows a binomial distribution and use a model that is specifically designed to deal with that distribution (and with several others). In particular, we would suggest fitting a GLIM, with binomial error and logit/probit link. We cannot give details here (see e.g. McCullagh & Nelder, 1989), but we only mention that this GLIM itself takes care of the characteristics of our dataset (non-normality, heteroscedasticity, non-linearity/non-additivity) and we may estimate parameters, test hypothesis and compare treatments, as we would normally do in the ANOVA/regression, but with far fewer problems in terms of distributional assumptions.

Such an approach has become popular for the analysis of quantal pesticide bioassays (logistic regression or probit regression; Finney, 1979) and it has been one of the first examples of GLIMs. However, those who are already familiar with GLIMs and intend to use them in replicated field trials (as in the above example) should not forget that, while distributional assumptions are relaxed, the independence assumption still holds. Indeed, observational units (the plants, in this example) should be randomly chosen and independent, while in replicated field trials they are generally clustered within experimental units (plots).

Analogously to subsampling (see above), this issue requires the use of a GLMM (see e.g. Bolker *et al.*, 2009), that includes the subject (plot/pot/Petri dish) as a random effect in the GLIM (Piepho, 1999). In other words, we treat pseudoreplication in GLIMs as we have shown in the case of linear models.

How to deal with multinomial counts

In some cases, the possible outcome of an experiment may not be binomial, but multinomial. An example is given by Mercer *et al.* (2006), who compared the seed status (germinated, dormant and dead) of a series of crop-wild sunflowers hybrids. In this case, traditional methods are difficult to apply, as the three counts for each variety, eventually transformed into proportions, should be used one at a time within an ANOVA.

With a GLIM, the seed status may be considered as a multinomial dependent categorical variable (as the above authors in fact did), by using a 'multinomial' family (an extension of the binomial distribution) and a logit link; the advantage with respect to a traditional approach should be clear. If the categories have a natural ordering, special kinds of GLIMs may be more appropriate (see next section).

Rating scale data

In several cases, experimental units are scored on a scale that does not carry any biological meaning, apart from arranging those subjects in a certain order. For example, herbicide phytotoxicity and efficacy are frequently scored on ordinal scales (e.g. from 1 to 9), by means of subjective visual observations. In some cases, the score itself is submitted to ANOVA, although this may not be appropriate, as statistical methods based on means and differences (such as the ANOVA) make sense only with metric data.

This situation is the same as that of multinomial data; indeed, we may (i) regard this kind of experiments as having nine possible outcomes (the nine phytotoxicity levels), (ii) count for each treatment the number of subjects (plots) in each of the nine categories and (iii) submit those counts to GLIM analysis (multinomial family and logit link). This is possible and we could even take advantage from the fact that categories are naturally ordered, by using some particular forms of GLIMs, such as the proportional odds model (a good example is shown again in Schabenberger & Pierce, 2002) and the 'threshold model' proposed by McCullagh and Nelder (1989). However, this requires a high number of replicates, while herbicide phytotoxicity field trials are generally carried out with no more than 3–4 replicates.

In these cases, non-parametric methods, such as rank-based methods, should at least be considered. Indeed, ordering observations from, say, highest to lowest, the rank for one observation is the number of observations higher or equal to that observation; such a rank is easily interpretable, as well as the difference in rank between two observations.

Examples of rank-based methods are the Kruskal-Wallis test that may correspond to the one-way ANOVA as a parametric test, and the Friedman test, that may correspond to the ANOVA for a randomised complete block design. Recent advances in this field have extended the range of applicability of rank-based methods beyond one-way layouts, up to factorial, split-plot and repeated measures designs (Shah & Madden, 2004).

Non-parametric statistics have been rather underrated by weed scientists, although rank-based methods have several good properties; they are easy to apply and understand and they are only slightly less powerful than their parametric equivalent, in case all the assumptions hold for the latter. A book-length treatment can be found in the classic work by Siegel and Castellan (1988) and in Brunner *et al.* (2001).

How to deal with counts with no natural denominator

In weed science, very frequently counts cannot be transformed into proportions, as there is no natural denominator. Examples are the number of weeds or seeds per square metre, the number of tillers or leaves per plant, the number of insects per leaf, etc.

These kinds of counts are traditionally submitted to ANOVA, either directly or after a square root or logarithmic transformation. This is acceptable, but the adoption of a GLIM may help reach more reliable conclusions by using the 'Poisson' error family with log link. Such an approach was used in an experiment relating to the counts of poppies on field plots, with five treatments and four blocks, as reported in Schabenberger and Pierce (2002). With quantitative explanatory variables, this approach is known as 'Poisson regression'.

Also in this case, possible problems with the clustering of observations within experimental units should be appropriately accounted for, e.g. by using a GLMM.

Time to event data

In weed science, researchers may be interested in time to event assessments, as in the case of germination studies. Frequently, the time course in the number of germinated seeds is analysed by using non-linear regression, even though this may pose problems related to non-normal error distribution, heteroscedasticity and serial correlation between the number of seeds counted in different dates on the same experimental unit (Petri dish, pot, plot). Furthermore, it is necessary to take into account viable seeds that have not yet germinated at the end of an experiment (censored observations).

These problems were reviewed by Scott *et al.* (1984), who proposed the use of survival analysis. This technique is normally used in medical research to model time to event data, in the presence of censored observations (Venables & Ripley, 2003); even though its real usefulness in weed science has yet to be elucidated, survival analysis may represent an option in some experimental situations. A good discussion of this is given by Scheiner and Gurevitch (1993).

GLIMs: things to care about when working with all kinds of counts

As mixed models, GLIMs are very powerful, even though they may represent a rather technical subject. The above-mentioned book of McCullagh and Nelder (1989) provides a complete reference, while Faraway (2006) gives a more application-oriented presentation and Molenberghs and Verbeke (2005) extend the problem to repeated measures.

For those who are already familiar with GLIMs, we would like to draw attention to two basic issues, strongly connected to each other:

1. Do not forget the clustering of observations within experimental units;
2. Check always for overdispersion.

Working with counts, observational units (plants, seeds, etc.) are almost always clustered within experimental units and so they are never truly independent; if we forget this, we unacceptably increase the assumed number of true replicates (i.e. we confuse true replicates with pseudoreplicates). We have already mentioned that in presence of clustered observations it is necessary to use a GLMM and include the observational unit as a random factor in the GLIM.

If not appropriately accounted for, the clustering of observations may result in the so-called 'overdispersion', i.e. a residual deviance that is much larger than would be expected if the model were correct. This problem is very frequent in GLIMs and, apart from clustered observations, it may arise for several other reasons (Faraway, 2006):

1. The selected family does not correspond to the real underlying distribution of the response variable. For example, if we use a Poisson family with weed counts, we assume that weeds are randomly distributed across fields, while their distribution is clearly aggregated (patchy).
2. Some important predictor is missing.
3. A few outliers are present.

In the presence of overdispersion, parameter estimates will still be reliable, but standard errors will be underestimated.

Overdispersion should always be checked: as a rule of thumb, we can suspect it whenever the residual deviance (or better the sum of squared Pearson residuals) is higher than its degrees of freedom. If we are in this situation, we should do one of the following:

1. Change the error family (e.g. in the frequent case of patchy distributed weeds, we might switch from Poisson to negative binomial);
2. Use a scale parameter (e.g. via a Pearson-type estimator), so that standard errors of estimates are appropriately adjusted;
3. In case of clustered observations, fit a GLMM (as mentioned above).

Other more advanced approaches have been suggested by Hughes and Madden (1995).

Aside from overdispersion, a further very important issue is that GLIMs may fail in reaching convergence and/or producing acceptable estimates with many zeros in count data. This situation may be common in weed surveys and it is not easy to deal with. One advice may be that if a treatment produces zeros only, it may be deleted to produce a reasonable analysis. Other suggestions might be given (Ming & Agresti, 2005), but this is one of the cases where, very frequently, we need to go back to ANOVA/regression, even though this method would seem less appropriate.

Several response variables: multivariate statistics

A great part of datasets collected by weed scientists are multivariate, in the sense that several variables are measured in each subject. More specifically, the weed flora (vegetation dataset of species abundances in sites or quadrats) in itself is a perfect example of a multivariate 'entity'. Very frequently, the different variables are isolated and analysed separately; in some cases this approach works well, but in other cases it does not permit insights into possible relationships among variables and key features of interest. As the consequence, multivariate analysis has been used in vegetation research since the 1950s, with the aim of exploring and summarising very complex datasets. Several examples may also be found in weed science. The subject is rather complex and a full treatment is far outside the aims of these paper. A review paper relating to multivariate methods in weed research is Kenkel *et al.* (2002). Another useful review (relating to microbiology) has been published by Ramette (2007) and a classical textbook is that of Legendre and Legendre (1998).

From an editorial perspective, the use of multivariate methods may raise some concern by referees, especially if they are not very experienced in the application of those methods. Some of the questions raised may be:

1. Is the use of multivariate analysis fully justified? As we mentioned at the beginning, a poor dataset may be easily hidden behind complex analyses.
2. Is the selected method appropriate to the aims? The world of multivariate analysis is multifaceted and some selection criteria appear to be rather arbitrary.
3. Are the assumptions for certain methods fully satisfied? Do not forget that also multivariate methods make several assumptions.
4. Are the effects significant or not? Multivariate methods frequently lack formal hypothesis testing, which may perplex the 'most traditional' weed scientists, but it may be fully justified to produce an exploratory analysis without significance testing.
5. Is the level of detail deep enough to be able to repeat the analysis? The use of multivariate methods may be preceded and followed by several optional types of data manipulation. For example, it is usually crucial to standardise different variables before submitting them to Principal Component Analysis (PCA) or similar methods.

Some of these questions may be trivial, but very frequently multivariate datasets are fed into available software, without having the necessary background to understand the output. We here give some advice, to avoid certain common pitfalls and misunderstandings, that may pose an obstacle to the process of paper evaluation and review.

From a general perspective, some details should never be neglected in the Materials and methods, that is:

1. Reasons for the choice of a particular multivariate method (the answer should logically follow from the aims of the experiment).
2. Clear identification of experimental units (objects) and variables (remember that experimental units are defined through the sampling process, before making measurements).
3. Clear indication of which pre-processing treatments (if any) for variables (recoding, standardisation, normalising transformation) have been performed. In some cases, this pre-processing is automatically performed by the software and authors should be aware of that.
4. Report clearly how missing data (if any) were handled.
5. Report whether basic assumptions were checked and how. Although violations of assumptions may not be very problematic, especially when formal hypothesis testing is not required, authors should always make this check and behave accordingly. In particular, keep in mind that normalising transformations used on one variable at a time do not necessarily imply that the overall distribution becomes multivariate normal.

6. Do not forget that the simultaneous use of differently distributed variables (continuous, ordinal, categorical, if appropriate to the selected method) and/or of variables of a different nature (agronomical, morphological, meteorological, biochemical, etc.) should always be very carefully and adequately motivated (Gower, 1971; Bramardi *et al.*, 2005). Keep also in mind that data with many zeros may pose a problem. For example, they may give rise to the so-called 'horseshoe' effect (Digby & Kempton, 1987).

Some more specific advice can be given with reference to the most important multivariate methods, organised according to their aims.

Describing data and reducing their dimensionality: ordination methods

Ordination methods are aimed at reducing the complexity of a multivariate dataset, with only a minor loss in its informative power. These methods are generally used in an exploratory setting and thus the lack of formal hypothesis testing should not be regarded as a problem; in this case the weed scientist's experience and knowledge of literature are the most important ingredients for a reliable interpretation of results.

The most widespread method of ordination is PCA, which is aimed at describing the variation in a set of original and correlated variables by using a new set of uncorrelated variables [principal components (PCs)], each obtained by a linear combination of the original ones. This method is very powerful and may give a very effective summarisation of complex datasets, although authors should not forget the following:

1. The PCA needs a set of quantitative variables, characterised by (i) a linear correlation structure, (ii) no outliers, (iii) no missing data and (iv) few zeros. Note that in weed surveys, the above criteria may not be met, especially when sites differ greatly for the composition of weed flora.
2. The PCA requires variables on comparable scales; otherwise, standardisation is required. Mention clearly if standardisation or other kind of 'pre-manipulation' has been performed.
3. Mention the percentage variance explained by each PC.
4. Mention whether some sort of rescaling on PCs has been used. This is particularly important in those cases where results are summarised by using 'biplots', as the selection of one particular scaling option strongly affects the interpretation of the plot, in terms of distances and angles. For example, one scaling may allow Euclidean distances between objects to be interpreted, but not those between variables, while another scaling does not permit either. A typical mistake is to drag and pull a PCA plot so it fits the layout, but to forget that axes need to be equally scaled for any geometric interpretations of distances and angles. A good reference for these issues is Digby and Kempton (1987).
5. Indicate the quality of representation of individuals and variables in the represented plane.
6. A minimum spanning tree (Gower & Ross, 1969) may be superimposed on the plane to help the interpretation of the relationships among individuals.

Giving more information is far beyond the objectives of this work, but it may be useful to mention other ordination techniques for those cases where PCA is not applicable. In particular, we mention Principal Co-ordinate Analysis (PCoA; also know as Metric Multidimensional Scaling; see Demey *et al.*, 2008; Gower, 1966), Correspondence Analysis and Non-Metric Multidimensional Scaling. All these methods have been reviewed in Kenkel *et al.* (2002); their selection is not easy and thus it is imperative that authors state clearly the reasons behind their choice.

Grouping observations – cluster analysis: classification methods

The term 'cluster analysis' embraces a wide range of techniques aimed at taking a certain number of individuals and discovering groups (clusters) of relatively similar individuals, based on a group of clustering variables. The analysis may be based on one of several (dis)similarity indices and can be performed by using a high number of different clustering techniques: we only mention that, among agglomerative clustering methods, we can count at least 10 algorithms (single linkage, complete linkage, group average, UPGMA, Ward's, just to name a few).

We will not go into detail here, but we would like to point out that cluster analysis is a very powerful technique and it is relatively easy to apply. Thus, it is very commonly used in several branches of plant science, for example molecular biology and genetics, and, sometimes, it is considered more 'exact' or 'precise' than ordination techniques (e.g. PCA), that, by retaining a reduced set of new variables (or components), result in a loss of information. Note, however, that classification methods also produce a loss of information when forming the clusters, because all agglomerative clustering algorithms distort the relationships between individuals by changing the definition of the original distances (or similarities) by the choice of the aggregation method (see Everitt, 1979 for more detail). We recommend authors to be very careful about:

1. Proper selection of distance/aggregation/(dis)similarity index (see e.g. Kosman & Leonard, 2005);
2. Proper selection of clustering method;
3. Proper determination of optimal number of clusters.

Furthermore, if one wants to go beyond a purely descriptive analysis, it is necessary to:

1. Use some kind of formal hypothesis testing procedure (Mohammadi & Prasanna, 2003), to assess the optimal number of clusters;
2. Assess the precision of the cluster, which is the stability of the formed branches if the experiment would be repeated. A bootstrap approach (Jain & Moreau, 1987) and/or the relation between the original and final distance matrices (Sokal & Rohlf, 1962) may be useful for hypothesis testing.

Furthermore, it should not be forgotten that ordination methods (like PCA) may be better than cluster analysis at discovering important groupings in the dataset and, above all, they may be better in showing when a real grouping does not exist (Venables & Ripley, 2003). Indeed, by its very construction, cluster analysis and dendrograms may be suggestive of groupings when no such groupings truly exist.

Individuals and/or variables partitioned in groups – canonical tools

The above methods work with independent individuals, without any natural intrinsic partitioning. In some cases, this natural partitioning exists: think for example about an experimental trial aimed at comparing the composition of weed flora with different agricultural systems, based on observations repeated in several years. Here, we have groups of data for each agricultural system and Multivariate Analysis of Variance (MANOVA), Canonical Discriminant Analysis and Canonical Variate Analysis may help determine: (i) whether the groups are statistically different from one another, (ii) which of the measured variables can better contribute to discriminate among groups and (iii) assign new observations to one of the groups.

In contrast to the previous multivariate methods that are mostly descriptive or exploratory, some formal hypothesis testing can be required here and thus it is very important that the dataset meets the following basic assumptions:

1. Joint distribution of variables should be approximately multivariate normal;
2. No outliers are present;
3. Within-group covariance matrix must be homogeneous (analogous to the assumption of homoscedasticity in ANOVA);
4. As in PCA, the underlying data structure must be linear;
5. The total number of sampling units must exceed the number of measured variables.

Authors should check all the above and, if necessary, take the appropriate correcting measures.

In other cases, variables are partitioned in two sets (such as one response set and one factor set; for example a set of weed species and a set of environmental variables) and we aim at determining the correspondence between them. Techniques such as CANonical CORrelation Analysis (CANCOR) may be appropriate. Roughly speaking, CANCOR may be seen as an extension to multivariate regression, where we have more than one response variable. The aim will be to find linear combinations of one set of variables that best correlate with linear combinations of the other set of variables (Everitt, 2005). Statistical tests are available to determine the significance of correlations, but they make the usual assumptions of multivariate linearity and normality, which should always be checked at the beginning.

Other canonical methods exist, that allow one to incorporate environmental covariate information in the definition of scores for objects and variables. We only mention Redundancy Analysis and Canonical Correspondence Analysis and refer to the above cited works for more detail.

Presenting the data

Based on experience, it is possible to make some simple suggestions on the appropriate presentation of statistics in *Weed Research*. Of course, they should not be taken too literally, as there are alternative correct ways to do things.

Description of experimental methods

1. The experimental design must be stated explicitly, especially with regard to replication, structure (e.g. factorial; nested; dose response) and layout (e.g. completely randomised; randomised complete block; Latin square; lattice). Where a split-plot design has been used, it must be clear which treatments were on the main plots and which were on the sub-plots and, for each stratum, it should be clear what layout has been chosen (e.g. main plots may be randomised according to a lattice design).
2. The type of analysis should be stated explicitly. For example, if the results are analysed by ANOVA, followed by a comparison of means using the LSD, the text must say so, rather than leaving it to the reader to assume this.

3. Where statistical procedures are generic for all experiments being reported, it may often be helpful for the details of the analyses to be placed under a separate sub-heading.
4. If an unconventional method of analysis is used, a source reference should be given.
5. It may be useful to refer to the computer package used, unless the statistical method is a standard one and it is clearly stated (clear explanation of the relationships between the factors and the type of effects).
6. If parametric methods were used, state clearly how basic assumptions were checked. This applies also to multivariate methods.
7. If a transformation of data was used, it should be stated what the transformation was and to which data sets it was applied (if not to all data). When logarithms were used, it should be clear whether they were natural (ln) or to base 10 (log), or any other base.
8. If problems with the basic assumptions can be suspected and no correcting measures were taken, reasons for this should be stated clearly.
9. With linear and non-linear regression, mention if lack-of-fit has been checked and how.
10. Report how any missing data or outliers were treated in the analysis.
11. When using GLIMs, state whether problems with overdispersion were observed and how they were accounted for.
12. When using exploratory multivariate methods such as PCA, PCoA or cluster analysis, clearly state how data were pre-processed and how (standardisation, transformation).
13. When producing graphical plots, such as biplots for PCA and similar techniques, explain exactly how scores in the plot were scaled and which geometrical interpretations are permissible in light of the scaling used. Also make sure that principal axes are equally scaled, for otherwise geometrical properties such as distances and angles cannot be reasonably interpreted.

Discussion of results

1. Avoid duplication of data in text and table or graph.
2. If you have to present a small amount of data, consider including them in the text.
3. Use graphs to show trends: with more than three levels a graphical representation will always convey far more meaning than a tabular one.
4. Dispersion plots and 'join the dots' can be used to emphasise trends in the data.
5. Every estimate (in text, tables and graphs) should be followed by a measure of variability (see later).
6. If curve fitting has been used, the observed means at each level should always be shown along with the regression line. In case of replicates, show only their means.
7. Standard errors of the regression parameters should be given, along with the equation. The value of R^2 is not essential where the predictor (x-axis) is an experimentally controlled variable.
8. When using ANOVA on a factorial set of treatments, the significance of interactions should be discussed explicitly in the text; 'main effects' should in general only be examined if there are no significant interactions.
9. Full ANOVA tables, showing residual mean squares, variance ratios etc., are not normally required. However, where analyses are complex and various contrasts are embedded, it may be helpful for the description of results to show the ANOVA table, but giving only the degrees of freedom and probability values.
10. Whenever MCP have been properly applied, letters may be used to display significance of mean comparisons in tables or graphs, but they should not make the table/graph less readable. Letters displays are generated by most packages only when the design is variance balanced, but there are procedures for generating letters that work also for unbalanced data (Piepho, 2004).
11. Do not allow discussion of which means do and do not differ significantly to obscure the message being conveyed. Limit statements about significance to those which have a direct bearing on the aims of the research.
12. If the objective of a study is to estimate the *size* of an effect, then that estimate should be stated explicitly, with confidence intervals, and not just that 'the difference was significant', or non-significant, as the case may be.
13. When regression analysis was used, base the discussion on an overall consideration of the whole regression trend [e.g. 'response to treatment increased with application date, although further increases were small after about 60 days after sowing' (DAS)] and not on pairwise comparisons of individual means (e.g. '20 DAS was significantly greater than 40 and 60 DAS; 80 DAS differed ($P \leq 0.05$) from 60 DAS, but not from 20 DAS or 40 DAS').
14. Arithmetic means in unbalanced factorial designs, ANCOVAS or mixed models could be misleading and should not be used when reporting results or comparing means. Instead, use least squares (or adjusted) estimates of the means.

Measures of data variability

We have already mentioned that the fundamental reason for statistical analyses is that experimental and survey data are variable. To describe these data adequately, it is therefore necessary to present each estimated value together with a measure of variability in text, tables or graphs.

In the frequent case of normally distributed datasets, the measure of variability should be either the standard deviation (SD), or the standard error of the estimate (SE; estimates include the mean, the regression coefficients, etc.). Sometimes, other measures are reported, such as the standard error of a difference (SED), the confidence interval for means or the LSD. All those measures carry a different meaning and they should be used appropriately, according to the information that is necessary to convey. Some suggestions may be given as follows.

1. Whatever measure you use, explicitly mention what it is and, if at all possible, be consistent throughout the paper.
2. Use the SD if you want to express the variability of a cohort of measures, with respect to the mean. Within these terms, the SD is purely a descriptive indicator.
3. Use the SE if you want to express the likely variability of the estimates that could be obtained by repeatedly collecting samples from the concerned population. Within these terms, SEs are inferential indicators, and they are definitely more important than SDs, if one intends to express the general value of the results, beyond the observed sample. On the other hand, however, the SE is not a suitable measure of dispersion of individual observations, as it is always smaller than the SD.
4. Use the SED, along with its number of degrees of freedom, in the case of comparisons of an unstructured set of treatments in a balanced experiment. When data are unbalanced (and thus the SED is not unique), you can report a mean value for the SED, together with some indication of the variability of the pairwise SEDs, such as their minimum and maximum.
5. The SED may be replaced by a critical difference (always with degrees of freedom and number of replicates); good choices are the LSD for CWE rate and Tukey HSD for FWE rate.
6. Report the confidence interval to express a range of possible values for some characteristic of a population. In this case, the use of SE is inappropriate, as it represents only a 68% confidence interval.
7. With non-normal data or when outliers are present, the above indicators may be deceptive indexes of variability. In these cases, more robust indicators might be preferred, such as the median, instead of the mean, the first and third quartile, instead of standard deviation.
8. In a balanced design, the standard error of the means for the different levels of the factor is constant because it is assumed that the population variance is also constant. Do not use the estimates of the SE calculated within each level of the factor. For an unbalanced design, the SE is obtained from the corresponding error term and it is given by most of the standard statistical computer packages.

Correct use of cut-off levels for P*-values. How to deal with (marginally) non-significant results?*

Reporting cut-off levels may give rise to some errors (see below), that should be avoided. It may be necessary to recall that statistical tests are done relative to some predetermined cut-off level α and significance is determined by looking at the real observed probability value P. If the test is significant, the observed P is $\leq \alpha$ (and not $P = \alpha$, which is nearly impossible), while if the test is non-significant, the observed P is $> \alpha$.

For the above reasons, wordings like: (i) differences were significant at $P = 0.05$; (ii) no significant differences were detected ($P = 0.05$); (iii) the LSD was 13 ($P \leq 0.05$), are all wrong and should not be used.

The situation becomes particularly tricky when the observed P is little above α (e.g. $P = 0.051$). Indeed, cut-off levels of confidence are purely arbitrary and they are directly analogous to levels of 'reasonable doubt' in a court case. If $\alpha = 0.05$, it is not very logical to say that one effect is real because $P = 0.049$ and another is not real because $P = 0.051$. Such an obsession with P values has been justly criticised by Goodman (1999), who emphasises that statistical significance should never be disjointed from biological significance. Indeed, the latter is mainly concerned with the real size of an effect: a large effect can be statistically non-significant, but biologically relevant, while a small effect may be statistically significant, but biologically uninteresting. Further references on this topic include Hilborn and Mangel (1997) and the website: http://www.tufts.edu/~gdallal/LHSP.HTM.

As a consequence, there may be some situations that require an adequate amount of caution. Whenever a statistical test is not significant compared with an arbitrary cut-off level, three things may be happening in reality: (i) there is no effect; (ii) there is a small effect, which cannot be distinguished from background noise; (iii) there is a big effect, but background noise is also very big. This latter aspect is particularly interesting in some biological disciplines, where the variability of data is intrinsically very high. Some advice can be given to deal with the above situations.

1. The cut-off levels of 0.05, 0.01 and 0.001 are conventionally used and accepted, but there is no logical reason why a cut-off probability level of 10% (P = 0.1) should not be used, whenever a lower degree of protection is needed to support a biologically relevant effect.
2. Report a precise P-value. This conveys far more information and permits the reader to reach his/her individual conclusions: reporting P = 0.053 or 0.53 instead of 'NS' makes a great difference (Marini, 1999).
3. Report the effect size together with confidence intervals; this may help readers to assess the biological relevance of an effect, beyond its statistical significance (Colegrave & Ruxton, 2003).

In any case, it is preferable to conclude that 'a difference could not be detected' rather than that 'there was no effect'. This was nicely put in a paper by Altman and Bland (1995) who expressed concern about the consequences of considering the 'absence of evidence' as an 'evidence of absence', when issues of public health are concerned. In weed science, such an issue is very relevant to recent research on genetically modified herbicide-tolerant crops (Perry *et al*., 2003).

Presenting transformed data

Unfortunately, when data have been transformed, the presentation of results can become messy. This is unfortunate, but we must not allow statistical rigour to be disregarded in favour of convenience. For example, it is not acceptable to state that analyses were performed on transformed data, but then present means and SE of non-transformed data 'for clarity'.

The main problem with transformed data stems from the fact that the measure of variability cannot be easily back-transformed. Possible approaches to this problem (each with advantages and drawbacks) may be:

1. Present only back-transformed means and back-transformed measures of variability, via the delta-method (Weisberg, 2005). This is the clearest approach, as both means and measures of variability are given in their original measurement unit, even though not all the statistical packages make the delta-method readily available. One should not forget that naively back-transformed means predict medians, not 'means' in the sense of expected values, which is not a disadvantage, however. Indeed, medians make more sense than means as measures of central tendency, in case of skewed (non-normal) data, which is almost always the case if a transformation has been found necessary to achieve normality.
2. Present only back-transformed means, but with letters to indicate the results of the significance test on the transformed data. This is certainly easier than using the delta-method, but the consequence is that no explicit measure of variability is given to the reader.
3. Present only the transformed means with their measure of variability. The drawback here is that their magnitudes will be meaningless to the reader.
4. In tables, present transformed means in parentheses alongside their back-transformed values, with the variability of the transformed data underneath (again in parentheses). The drawback is that this may result in an excessively cluttered table. However, cluttering can often be minimised by reducing the amount of uninformative data being presented (authors often tend to present all their data, rather than just those directly relevant to their objective), or by having a greater number of smaller tables.
5. In figures, means can be shown on transformed axes, but with the axes labelled according to back-transformed values. For example, the 'tick marks' on a $\log_{10}$-transformed axis can be shown as 0.1, 1, 10, 100, 1000, etc. The SE of the transformed data can then be depicted by a single error bar.

Final remarks

The references that we refer to are of course but a small sample of an extensive literature. As well as regularly scanning *Weed Research* we also encourage readers to browse both old and new statistics books.

One of the biggest concerns in writing statistical guidelines is that they may be interpreted too literally, which may prevent some good research from being successfully published. Therefore, we would like to reinforce the idea that statistical methods are not a set of recipes whose mindless application is required by convention; each experiment or study may involve subtleties that guidelines cannot cover. Scientists and editors are therefore warned: suggestions should not be taken as fixed rules!

The only two rules that we can reasonably give in a fairly prescriptive manner are the following:

1. *State clearly the objective of the experiment*. The choice of statistical methods depends always on the aims, which should be stated very early (preferably at the end of *Introduction*). All design, analysis and interpretation should then flow on logically from this point.
2. *If in doubt, consult a biometrician*. For weed scientists, statistical training may not be at the same level as biological, chemical or agricultural training. If you are even slightly unsure about a statistical design,

method or about the way in which a result has been interpreted, consult a biometrician.

With respect to this latter point, we must say, from experience, that too often a biometrician is consulted only at the moment of data analysis. Statistical help at this late stage does not necessarily guarantee the statistical validity of the paper. Indeed, a biometrician might help also in designing the experiment and also during the writing process, to make sure that methods are properly described and conclusions are supported by the data (Fenlon, 1995). Apart from the above rules, do not let common sense and clear thinking be replaced by the rigid application of statistical orthodoxy.

References

Abdi H (2007) The Bonferroni and Sidak corrections for multiple comparisons. In: *Encyclopedia of Measurement and Statistics* (ed. N Salkind), 103–107. Sage publication, Thousand Oaks, CA, USA.

Ahrens WH, Cox DJ & Budwar G (1990) Use of the arcsin and square root transformation for subjectively determined percentage data. *Weed Science* **38,** 452–458.

Altman DG & Bland JM (1995) Statistics notes: absence of evidence is not evidence of absence. *British Medical Journal* **311,** 485.

Anonymous (1988) Guidance for the use and presentation of statistics in Weed Research. *Weed Research* **28,** 139–144.

Atkinson AC (1985) *Plots, Transformations and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis*. Oxford University Press, Oxford, UK.

Bolker BM, Brooks ME, Clark CJ *et al.* (2009) Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution* **24,** 127–135.

Box GEP & Cox DR (1964) An analysis of transformations. *Journal of the Royal Statistical Society* **B-26,** 211–252.

Bramardi SJ, Bernet GP, Asins MJ & Carbonell EA (2005) Simultaneous agronomic and molecular characterization of genotypes via the generalized procrustes analysis: an application to cucumber. *Crop Science* **45,** 1603–1609.

Brunner E, Domhoff S & Langer F (2001) *Nonparametric Analysis of Longitudinal Data in Factorial Experiments*. John Wiley and Sons, New York, USA.

Carroll RJ & Ruppert D (1988) *Transformation and Weighting in Regression*. Chapman and Hall, London, UK.

Clayton MK (2007) How should we achieve high-quality reporting of statistics in scientific journals? A commentary on "Guidelines for reporting statistics in journals published by the American Physiological Society". *Advances in Physiology Education* **31,** 302–304.

Cochran WG & Cox GM (1957) *Experimental Designs*, 2nd edn. John Wiley & Sons, New York, USA.

Colegrave N & Ruxton GD (2003) Confidence intervals are a more useful complement to nonsignificant tests than are power calculations. *Behavioural Ecology* **14,** 446–447.

Cousens R (1988) Misinterpretetion of results in weed research through inappropriate use of statistics. *Weed Research* **28,** 281–289.

Demey J, Vicente-Villardon J, Galindo-Villardon M & Zambrano A (2008) Identifying molecular markers associated with classification of genotypes by external logistic biplots. *Bioinformatics* **24,** 2832–2838.

Digby PGN & Kempton RA (1987) *Multivariate Analysis of Ecological Communities*. Chapman and Hall, London, UK.

Dormann CF, McPherson JM, Araujo MB *et al.* (2007) Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography* **30,** 609–628.

Draper NR & Smith H (1998) *Applied Regression Analysis*, 3rd edn. John Wiley & Sons, New York, USA.

Edward D & Berry JJ (1987) The efficiency of simulation-based multiple comparisons. *Biometrics* **43,** 913–928.

Everitt B (1979) Unresolved problems in cluster analysis. *Biometrics* **35,** 169–181.

Everitt B (2005) *An R and S-PLUS Companion to Multivariate Analysis*. Springer-Verlag, London, UK.

Faraway JJ (2004) *Linear Models with R*. Chapman & Hall/CRC, Boca Raton, FL, USA.

Faraway JJ (2006) *Extending the Linear Model with R. Generalized Linear, Mixed Effects and Nonparametric Regression Models*. Chapman & Hall, Boca Raton, FL, USA.

Fenlon JS (1995) How can a statistician improve your interpretation? *Pesticide Science* **45,** 77–82.

Finney D (1979) Bioassay and the practice of statistical inference. *International Statistical Review* **47,** 1–12.

Garrett KA, Madden LV, Hughes G & Pfender WF (2004) New applications of statistical tools in plant pathology. *Phytopathology* **94,** 999–1003.

Goodman SN (1999) Toward evidence-based medical statistics. 1: The *P* value fallacy. *Annals of Internal Medicine* **130,** 995–1004.

Gower JC (1966) Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* **53,** 325–338.

Gower JC (1971) A general coefficient of similarity and some of its properties. *Biometrics* **27,** 857–874.

Gower JC & Ross GJS (1969) Minimum spanning trees and single linkage cluster analysis. *Applied Statistics* **18,** 54–64.

Hilborn R & Mangel M (1997) *The Ecological Detective – Confronting Models with Data*. Princeton University Press, Princeton, NJ, USA.

Hothorn T, Bretz F & Westfall P (2008) Simultaneous inference in general parametric models. *Biometrical Journal* **50,** 346–363.

Hsu JC (1996) *Multiple Comparisons. Theory and Methods*. Chapman & Hall, London, UK.

Hughes G & Madden LV (1995) Some methods allowing aggregated patterns of disease incidence in the analysis of data from designed experiments. *Plant Pathology* **44,** 927–943.

Jain A & Moreau J (1987) Bootstrap technique in cluster analysis. *Pattern Recognition* **20,** 547–568.

John JA & Williams ER (1995) *Cyclic and Computer Generated Designs*. Chapman & Hall, London, UK.

Johnson SB & Berger RD (1982) On the status of statistics in phytophatology. *Phytophatology* **72,** 1014–1015.

Kenkel NC, Derksen DA, Thomas AG & Watson PR (2002) Multivariate analysis in weed science research. *Weed Science* **50,** 281–292.

KOSMAN E & LEONARD KJ (2005) Similarity coefficients for molecular markers in studies of genetic relationships between individuals for haploid, diploid, and polyploid species. *Molecular Ecology* **14,** 415–424.

LECLERG EL, LEONARD WH & CLARK AG (1962) *Field Plot Technique*. Burgess Publishing Company, Minneapolis, MN, USA.

LEGENDRE P & LEGENDRE L (1998) *Numerical Ecology*. Elsevier, Amsterdam, the Netherlands.

LITTELL RC, MILLIKEN GA, STROUP WW, WOLFINGER RD & SCHABANBERGER O (2006) *SAS for Mixed Models*, 2nd edn. SAS Publishing, Cary, NC, USA.

LUMLEY T & HEAGERTY P (1999) Weighted empirical adaptive variance estimators for correlated data regression. *Journal of the Royal Statistical Society B* **61,** 459–477.

MARINI RP (1999) Are nonsignificant differences really not significant? *HortScience* **34,** 761–762.

MAXWELL SE & DELANEY HD (1990) *Designing Experiments and Analysing Data*. Wadsworth, Belmont, CA, USA.

MCCULLAGH P & NELDER JA (1989) *Generalized Linear Models*. Chapman & Hall, London, UK.

MEASURES M, WEINBERGER P & BAER H (1973) Variability of plant growth within controlled-environment chambers as related to temperature and light distribution. *Canadian Journal of Plant Science* **53,** 215–220.

MERCER KL, SHAW RG & WYSE DL (2006) Increased germination of diverse crop–wild hybrid sunflower seeds. *Ecological Applications* **16,** 845–854.

MILLIKEN GA & JOHNSON DE (1984) *Analysis of Messy Data: Designed Experiments*, Vol 1. Lifetime Learning Publ, Belmont, CA.

MING Y & AGRESTI A (2005) Random effect models for repeated measures of zero-inflated count data. *Statistical modeling* **5,** 1–19.

MOHAMMADI SA & PRASANNA BM (2003) Analysis of genetic diversity in crop plants–salient statistical tools and considerations. *Crop Science* **43,** 1235–1248.

MOLENBERGHS G & VERBEKE G (2005) *Models for Discrete Longitudinal Data*. Springer, New York, USA.

MORRISON DA & MORRIS EC (2000) Pseudoreplication in experimental designs for the manipulation of seed germination treatments. *Austral Ecology* **25,** 292–296.

NELDER J (1994) The statistics of linear models: back to basics. *Statistics and Computing* **4,** 221–234.

NIELSEN OK, RITZ C & STREIBIG JC (2004) Nonlinear mixed-model regression to analyze herbicide dose–response relationships. *Weed Technology* **18,** 30–37.

PEARCE SC (1992) Data analysis in agricultural experimentation. I. Contrasts of interest. *Experimental Agriculture* **28,** 245–253.

PERRY JN, ROTHERY P, CLARK SJ, HEARD MS & HAWES C (2003) Design, analysis and statistical power of the farm-scale evaluations of genetically modified herbicide-tolerant crops. *Journal of Applied Ecology* **40,** 17–31.

PHELPS K (1991) Some problems in the statistical analysis of trials for resistance screening. *Plant Pathology* **40,** 340–341.

PIEPHO HP (1994) Best linear unbiased prediction (BLUP) for regional yield trials: a comparison to additive main effects and multiplicative interaction (AMMI) analysis. *Theoretical Applied Genetics* **89,** 647–654.

PIEPHO HP (1997) Analysis of a randomized complete block design with unequal subclass numbers. *Agronomy Journal* **89,** 718–723.

PIEPHO HP (1999) Analysing disease incidence data from designed experiments by generalized linear mixed models. *Plant Pathology* **48,** 668–674.

PIEPHO HP (2003) The folded exponential transformation for proportions. *The Statistician* **52,** 575–589.

PIEPHO HP (2004) An algorithm for a letter-based representation of all-pairwise comparisons. *Journal of Computational and Graphical Statistics* **13,** 456–466.

PIEPHO HP, BÜCHSE A & RICHTER C (2004) A mixed modelling approach for randomized experiments with repeated measures. *Journal of Agronomy and Crop Science* **190,** 230–247.

PIEPHO HP, WILLIAMS ER & FLECK M (2006) A note on the analysis of designed experiments with complex treatment structure. *HortScience* **41,** 446–452.

PINHEIRO JC & BATES DM (2000) *Mixed-effects Models in S and S-Plus*. Springer-Verlag, New York, USA.

PLANT RE (2007) Comparison of means of spatial data in unreplicated field trials. *Agronomy Journal* **99,** 481–488.

RAMETTE A (2007) Multivariate analysis in microbial ecology. *FEMS Microbiological Ecology* **62,** 142–160.

RITZ C & STREIBIG JC (2008) *Nonlinear Regression with R*. Springer-Verlag, New York, USA.

SCHABENBERGER O & PIERCE FJ (2002) *Contemporary Statistical Models for the Plant and Soil Sciences*. Taylor & Francis, CRC Press, Boca Raton, FL, USA.

SCHABENBERGER O, THARP BE, KELLS JJ & PENNER D (1999) Statistical tests for hormesis and effective dosages in herbicide dose response. *Agronomy Journal* **91,** 713–721.

SCHEINER SM & GUREVITCH J (1993) *Design and Analysis of Ecological Experiments*, 1st edn. Oxford University Press, Oxford, UK.

SCOTT SJ, JONES RA & WILLIAMS WA (1984) Review of data analysis methods for seed germination. *Crop Science* **24,** 1192–1199.

SEARLE SR (1987) *Linear Models for Unbalanced Data*. John Wiley and Sons, New York, USA.

SHAH DA & MADDEN LV (2004) Nonparametric analysis of ordinal data in designed factorial experiments. *Phytopathology* **94,** 33.

SHAW RG & MITCHELL-OLDS T (1993) ANOVA for unbalanced data: an overview. *Ecology* **74,** 1638–1645.

SIEGEL S & CASTELLAN NJ (1988) *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, New York, USA.

SOKAL R & ROHLF R (1962) The comparison of dendrograms by objective methods. *Taxon* **11,** 33–40.

STREIBIG JC (1988) Herbicide bioassay. *Weed Research* **28,** 479–484.

STREIBIG JC, RUDEMO M & JENSEN JE (1993) *Dose–Response Curves and Statistical Methods*. CRC Press, Boca Raton, FL, USA.

VENABLES WN & RIPLEY BD (2003) *Modern Applied Statistics with S. Statistics and Computing*. Springer-Verlag, New York, USA.

WEISBERG S (2005) *Applied Linear Regression*. John Wiley & Sons Inc, New York, USA.

WHITTINGHAM MJ, STEPHENS PA, BRADBURY RB & FRECKLETON RP (2006) Why do we still use stepwise modelling in ecology and behaviour? *Journal of Animal Ecology* **75,** 1182–1189.