ISSUES CONCERNING THE DESIGN AND ANALYSIS OF COMPARATIVE FIELD EXPERIMENTS

Sue J. Welham & Suzanne J. Clark outlines the value of appropriate statistics in biological assays
Biomathematics and Bioinformatics Division, Rothamsted Research, Harpenden AL5 2JQ, UK

Keywords: analysis of variance, blocking, contrasts, data transformation, experimental design, multiple comparison test, randomization, replication.

Introduction

In scientific publications, a clear report of an appropriate statistical analysis allows independent assessment of the results and adds weight to the conclusions. Poorly described statistical analyses of field experiments give little justification for the conclusions drawn. However, to gain maximum benefit, statistical analysis should always be considered at the planning stage of the experiment, in the choice of treatments applied and the experimental design. Good choices at this stage can enable a suitable statistical analysis and reduce uncertainty in treatment estimates. This paper gives guidelines on the issues that should be considered at the design and analysis stages to facilitate good experimentation and reporting. These are principles and guidelines rather than rules, as the best choices for any particular experiment will depend on the context and aims of the experiment. Further general information on the design and analysis of field crop experiments aimed at practitioners can be found in Steel & Torrie (1990), Mead, Curnow & Hasted (1993) or Snedecor & Cochran (1989). However, these texts may be less helpful to scientists unused to mathematical notation. A better solution may be an early approach for advice to a statistical consultant with experience in the design and analysis of agricultural field plot experiments.

Choice of treatments

The treatments to be applied to field experiments must be chosen to specifically address the aims of the experiment. In the simplest case, one *treatment factor* (e.g. cultivar or herbicide regime) may be tested at each of several *levels*. Treatment factor levels may be *qualitative*, e.g. different cultivars, or *quantitative* with an underlying numeric value, e.g. rate of application of pesticide or fertiliser. All other factors that may affect the results should be held constant (or as constant as possible) across the experiment. Alternatively, two or more factors may be simultaneously tested with all combinations of the levels of the different factors present in the trial - such an experiment is said to have a *factorial* treatment structure. The general advantage of a factorial structure is that, because all combinations are present, it gives an interpretable test for overall differences within each of the *main effects* and also for their *interaction*. The main effect of a factor comprises an overall comparison between the levels of that factor averaged over all levels of the other factors applied. An interaction occurs if the response across levels of one factor differs according to the levels of other factors applied. Interactions can often be most clearly seen as non-parallel lines in graphical plots, demonstrated for a 2 × 2 factorial structure (two treatments, each with two levels) in Figure 1.

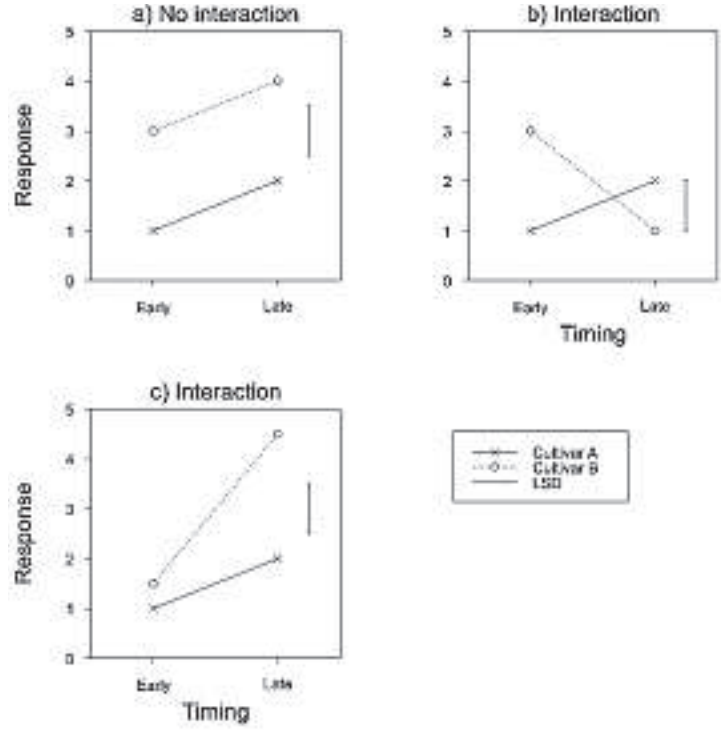

Figure 1. **Possible responses to treatments in a field experiment with a 2 × 2 factorial treatment structure consisting of two fungicide timings (early or late) for two cultivars (A or B): no interaction between treatment factors (a) or significant interactions (b,c).**

To illustrate these concepts, consider a factorial experiment done at Rothamsted in the 1997/98 growing season to study the effect of fungicide regime on development of light leaf spot disease on winter oilseed rape cultivars (Steed *et al.*, 1999). Ten different fungicide timings were used, including two control treatments: a negative control (no fungicide applied) to give a severe epidemic, and a

DOI: 10.1564/17aug10

positive control (routine monthly fungicide application October-March) to give maximum disease control possible with the fungicide. The two cultivars used were Capitol (resistant to light leaf spot) and Bristol (susceptible). There were thus two treatment factors: fungicide regime (ten levels) and cultivar (two levels), and all 20 combinations of the two factors were used. It was hence possible to test both the main effects of fungicide regime and cultivar, and also their interaction, i.e. whether the effect of fungicide regime differs between cultivars.

In some cases, the simple factorial structure cannot always be applied, for example, if some treatment combinations are not scientifically valid or physically possible. In this case, the treatment structure should be designed so that specific *contrasts*, i.e. treatment comparisons, can be used to examine questions of interest.

Choice of design

Once a set of treatment factors and an experimental location has been chosen, then the experimental design can be constructed. The three crucial elements of experimental design are *replication*, *randomization* and *blocking*. In practice, the experimental design also needs to consider the scale at which experimental treatments can be applied, the homogeneity of the field, whether treatment levels applied to neighbouring plots may interfere with each other, and the replication required for each treatment combination.

The scale at which the different levels of each treatment factor can be applied will determine the size of the *experimental unit*. This may differ between treatments: for example, different cultivars may be sown in small plots, but machinery constraints may mean that fungicide can be applied only to larger areas. The experimental unit for each treatment factor is defined as the smallest unit of experimental material to which different levels are applied. In our example (Steed *et al.*, 1999), the experimental unit for a cultivar is a small plot, and the experimental unit for fungicide regime is a set of two small plots. Replication for each treatment is then the number of experimental units to which each level is applied. The structure of the experimental units for each treatment must be reflected in the statistical analysis (see Table 1 later).

Given the sets of experimental units, levels of each treatment factor must be assigned at random to each unit to give a valid experimental design. The simplest design is the *completely randomized* design, where the experimental unit for all treatments is the same (a plot), and treatment combinations are randomly allocated to plots. This design is rarely used for field experiments, however, as there is usually background heterogeneity within the experiment. A more efficient type of design involves grouping, or *blocking*, units that are expected to be reasonably homogenous, i.e. expected to give similar results under the same treatment. Blocking can account for variation between different areas of the field that would otherwise inflate the estimate of *residual error* (i.e. the between-unit or background variability, the portion of total variability unaccounted for by the treatments applied). Blocks may reflect field characteristics, e.g. differences in pH, drainage or soil-born disease, or fertility trends, or can be used to guard against the impact of outside factors, e.g. the spread of pests or diseases into the experiment from one (or two) edges of the field. It follows that the choice of blocks and block sizes should be independent of the choice of treatments. In practice a compromise is usually made: the optimal block size is not usually known, so a reasonable block size is chosen so that a standard design can be used. Introduction of blocking into the design means that variation may be considered within each different level of blocking, or *stratum*, of the design.

Randomized complete block designs, with each combination of treatments appearing once in each block, are often used for field experiments as this is the simplest design that allows for some blocking. *Latin square* designs are used where gradients may exist in two perpendicular directions across a field. *Split-plot* designs are appropriate where different treatments require experimental units of different sizes. The field layout for the split-plot winter oilseed rape experiment with 10 fungicide regimes and 2 cultivars (Steed *et al.*, 1999) is shown in Figure 2. This design has three replicates of each treatment combination, with different fungicide regimes applied to whole plots, and the two cultivars grown on sub-plots within each whole plot. Randomized layouts for these standard designs can be generated by most statistical packages. In each of these cases, these may not be the optimal designs if there is a large number of treatment combinations (so that block sizes become very large) or if the background is very heterogeneous. In these cases, *incomplete block* designs or *general row-column* designs that use small blocks, with only a subset of treatments within each block, may be more efficient (Cochran & Cox, 1957, Mead, 1988, or Whitaker *et al.,* 2002, give more details). With all designs, randomization is essential to ensure

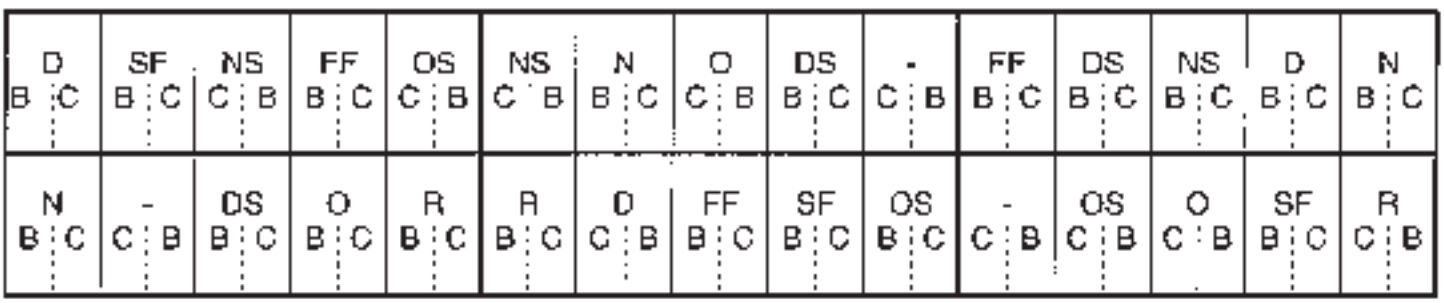

Figure 2: **Field layout of an experiment with a split-plot design to examine effects of 10 fungicide regimes [untreated control (-), October (O), October + March (OS), November (N), November + March (NS), December (D), December + March (DS), March + April (SF), April (FF), routine monthly October-March control (R)] on development of light leaf spot disease on winter oilseed rape cultivars Bristol (B) or Capitol (C) (Steed *et al.*, 1999).**

that treatment estimates are unbiased. A non-standard design may also be considered if neighbouring treatment combinations are expected to interfere with one another, for example, application of an effective fungicide may reduce disease inoculum received by neighbouring plots. Guard rows between plots are often sufficient to reduce these effects, but in extreme cases a restricted randomization may be used so that treatment combinations are allowed as neighbours only where interference is not expected (David & Kempton, 1996) or so that all neighbour combinations occur equally and balance out the interference effects (neighbour-balanced designs, e.g. Bailey, 1984).

The advantage of standard designs is that they naturally lead to a straightforward analysis. However, analyses for non-standard designs are now widely available (Piepho *et al,* 2003). The chosen balance between the safety net of a familiar analysis and potential gains in the efficiency in estimation of treatment differences may depend on the statistical knowledge and skill within the experimental team. Again, the inclusion of a statistical consultant within the team will help to ensure that an appropriate choice is made.

In all cases, a dummy analysis should be carried out on the proposed design to confirm that all treatment effects or contrasts of interest are estimable, and that they are estimated with reasonable relative precision. The dummy analysis should also confirm that sufficient residual *degrees of freedom* (generally recommended to be 12-20) are present (in the relevant strata, see Table 1) to allow the background variability to be estimated with some confidence. Where information on background variability is available from previous experiments, a preliminary analysis (a power study) can be done to determine the power of the design to detect a specified size of treatment difference. This can help to avoid the waste of time, effort and expense involved in doing an under-replicated experiment that cannot detect treatment differences of the size required, or an over-replicated experiment which gains no useful information for the extra resources used. For an under-replicated experiment, it may be better to decrease the number of treatment factors (or levels) tested in order to increase replication of the remainder.

Analysis

Subject to the checks described below, for standard designs the statistical analysis proceeds via the multi-stratum *analysis of variance* (ANOVA). The ANOVA table for the logit-transformed response variable %stem area affected by disease from the split-plot experiment shown in Figure 2 is given in Table 1. This analysis preserves the blocking structure of the experiment so that treatment terms (main effects or interactions) are assessed with respect to the appropriate estimate of background variation, provided by the residual mean squares. For each treatment term, the ratio of variation amongst treatments to the appropriate background variation indicates whether treatment differences could have occurred by chance, or if there is evidence that real treatment differences exist. This stratified analysis of variance is provided as default by the GenStat statistical system (Payne, 2003, Chapter 4) but has to be explicitly constructed in other packages. For non-standard unbalanced designs, the design can be expressed as a *mixed model*, with the blocking structure used as the random model, and the analysis done using the restricted maximum likelihood (REML) method (Piepho *et al.,* 2003) available in many standard statistical packages. In either case, any non-treatment factors that were not adequately controlled and subsequently found to vary across the experiment, and which might influence the results, can be included as covariates in the analysis.

Analysis of variance assumes that the data can be represented in terms of the block and treatment structures as an additive linear model, i.e. that each component of the structure adds (or subtracts) a consistent amount to the total response. For example, for a randomized complete block design with one treatment factor, the underlying model can be written as

$$y_{ij} = m + b_i + t_j + e_{ij}$$

i.e. the response y_{ij} for treatment j in block i can be represented as a sum of effects: the overall mean (m), the effect of the ith block (b_i), the effect of the jth level of the

Table 1: Analysis of variance table for analysis of light leaf spot severity data [logit (% stem area affected)] from a field experiment with a split-plot design with 10 fungicide regimes (applied to whole plots) and two winter oilseed rape cultivars (applied to sub-plots) (Steed *et al.*, 1999). All numbers rounded to 2 decimal places after calculation.

Source of variation	Degrees of freedom (d.f.)	Sums of squares (s.s.)	Mean square (m.s.)	Variance ratio (v.r.)	F-probability (p)
Block stratum	2	1.53	0.77	3.39	
Block.Whole-plot stratum					
Fungicide	9	14.48	1.61	7.13	<.001
Residual	18	4.06	0.23	1.41	
Block.Whole-plot.Sub-plot stratum					
Cultivar	1	21.70	21.70	135.50	<.001
Cultivar.Fungicide	9	1.82	0.20	1.26	0.314
Residual	20	3.20	0.16		
Total	59	46.80			

treatment (t_j) and the residual error (e_{ij}). If this model is implausible, then analysis of variance may not be helpful in interpreting the data. A special case occurs when the effects act multiplicatively rather than additively, then a logarithmic transform of the data may be expected to respond on an additive scale.

Analysis of variance also assumes that the errors on the response variable are independent, Normally distributed and have equal variances. These assumptions can be checked by considering properties of the data and by examining properties of the residuals from the analysis, as these are the best available estimates of the errors. Formal tests, such as Bartlett's test for equality of variances between treatment combinations, are available but graphical examination of the residuals is a useful preliminary step in detecting potential departures from the assumptions. A plot of residuals against fitted values will show any trend left in the residuals (which indicates the model may be inadequate) and whether the residual variance (indicated by spread about zero) is constant across the range of fitted values (to meet the assumption of equal variances). Differences in behaviour amongst treatment combinations may be detected if the points are distinguished accordingly using different colours or symbols. Where the residuals are compatible with a Normal distribution, then a plot of the ordered residuals against quantiles of the standard Normal distribution (usually called a Normal or Q-Q plot) should show an approximately straight line and a histogram of the residuals should have a symmetrical, bell-shaped distribution. However, for small data sets even genuine samples from the Normal distribution may show quite large deviations from the ideal shapes and patterns. Independence of the residuals may be investigated by plotting them according to the field layout: no pattern should be discernable. This assumption is commonly violated for repeated measurements, where several successive samples are taken as a time-course from each plot. Such measurements from the same plot may show unequal correlation across time, and the analysis should take account of this (Diggle *et al*, 1994).

The assumption that the errors are Normally distributed implies that the data are measured on a continuous scale, or on a close-to-continuous scale, without limits. Count data ($y \geq 0$) and percentage data ($0 \leq p \leq 100$) are obvious exceptions. For count data, variance often increases as the mean increases. In this case, a log-transformation ($z=\log(y)$, or $z=\log(y+1)$ if zero counts are present) maps onto an unlimited scale and may give an approximately equal variance over the range of the data. However, this transformation may be inadequate if there are many zero counts.

Percentage data are often calculated from the incidence (x) within a sample of size n from each plot as $p=100x/n$. Use of the logit transformation ($z=\log(x/(n-x))$, or $z=\log((x+1)/(n+1-x))$ if incidences of 0 or n are present) may be appropriate for such data. Where data are measured directly as a percentage, e.g. disease severity as percentage leaf area affected, then a nominal value of $n=100$ can be used in the transformation. A logit transformation (with $n=100$) was used on data for % oilseed rape stem area affected by light leaf spot before analysis (Table 1), and a set of residual plots is shown in Figure 3. These plots are reasonable given the sample size: although there is a possible trend in the plot of residuals against fitted values, no other variable was related to this trend and the variance was reasonably constant over the range of fitted values. The logit transformation may be inadequate if there are many values at the limits of the percentage scale (0 or 100) or where the percentages have been calculated as incidence from a small sample, e.g. the number of affected plants out of a sample of only ten plants per plot. Larger samples are therefore preferred for measurement of incidence. For either count or incidence data, if the transformation is unsuccessful in generating data that is suitable for analysis of variance, then a generalized linear model (GLM) with an appropriate error distribution (e.g. Poisson distribution for counts or binomial distribution for proportions) should be used (e.g. Payne, 2003, Chapter 3), with care taken to incorporate the blocking structure of the experiment. Scores that summarise an underlying scale should generally be avoided, and the underlying scale should be used explicitly wherever possible. Scores often do not correspond to a linear scale, so that an "average score" is not interpretable, and in this case the analysis of variance model given above is usually inappropriate.

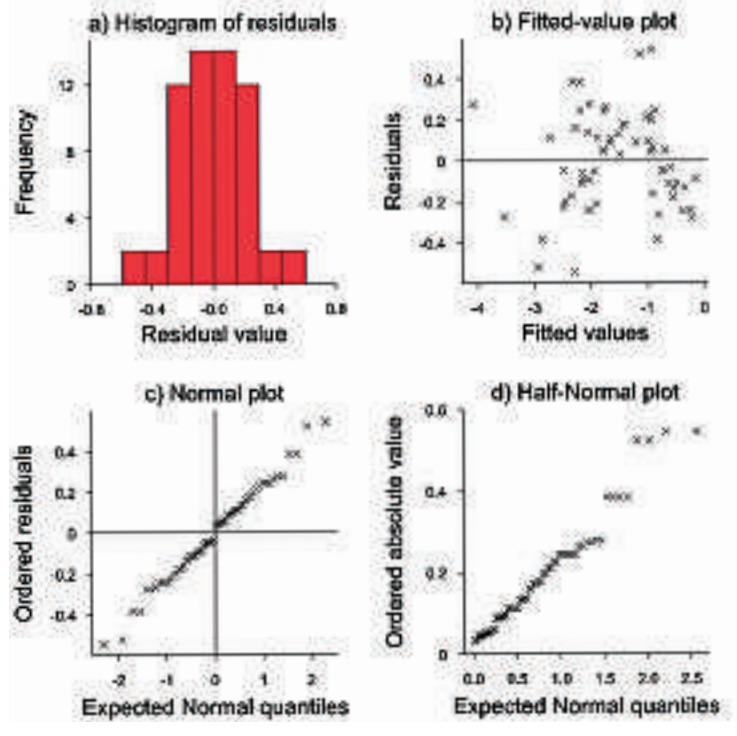

Figure 3: **Residual plots for logit (% winter oilseed rape stem area affected with light leaf spot) from split-plot experiment (Steed *et al.*, 1999) analysed in Table 1, used to check assumptions of analysis of variance: (a) histogram of residuals should show symmetric distribution; (b) plot of residuals against fitted values should show no trend and constant variation about zero; (c,d) Normal and half-Normal plot of ordered residuals or absolute value of residuals against the expected Normal quantiles should show approximately straight lines.**

Interpretation and presentation of results

The analysis of variable table (e.g. Table 1) is used to produce an F–statistic for each treatment term (the ratio of the variation due to the treatment term relative to the appropriate residual variation). An F-statistic is used to test the null hypothesis that all treatment effects within the term are equal (i.e. zero) against the alternative hypothesis of differences amongst treatment effects. The F-statistics can therefore be used to screen treatment main effects and interactions to detect terms where differences exist. Under the null hypothesis (no differences within the term) the statistic is distributed as an F-distribution with two component degrees of freedom: the first equal to the degrees of freedom of the treatment term and the second equal to the residual degrees of freedom of the stratum where the term is estimated. If the F-statistic is so large that a value of that size is unlikely to occur if the null hypothesis is true, this is taken as evidence against the null hypothesis. The p-value for the test is the probability of an F-statistic of that size (or larger) occurring if the null hypothesis is true. In analysis of data from an experiment with a factorial treatment structure, the highest-order interactions (e.g. cultivar × fungicide in Table 1) should be examined first. If these are significant (the threshold usually being taken as $p \leq 0.05$), then the model cannot be simplified further because the behaviour of one treatment factor changes depending on the levels of another. If these interactions are not significant, then lower-order interactions (and finally main effects) can be examined. This iterative procedure is followed until the model cannot be simplified further. The analysis in Table 1 indicates no evidence of interaction between cultivar and fungicide regime ($p=0.314$), but strong evidence of differences amongst cultivars and amongst fungicide regimes ($p<0.001$).

The aim is then to estimate (quantify) the effect of different treatment factors. Tables of means should be produced for the significant terms: for each treatment factor a table should be produced for the highest-order significant interaction that includes that factor. Tables of main effects should therefore be examined only if a factor shows no interaction with all other factors: if an interaction is present, then the main effect table does not produce meaningful predictions. Tables of means should be produced with standard errors of differences (SEDs) or least significant differences (LSDs) that can be used to evaluate specific differences of interest. In both cases, the degrees of freedom (d.f.) associated with the SED or LSD should be specified: this is the residual d.f. from the stratum in which the treatment mean was estimated. However, the computation of many pair-wise comparisons should be avoided, see below. Pre-defined contrasts of interest can be incorporated into the analysis of variance by partitioning the total sum of squares and treatment degrees of freedom appropriately. For treatment factors with an underlying numerical scale (e.g. fungicide dose) the form of the response across the numeric scale can be examined graphically by plotting treatment means (y-axis) against the numeric scale (x-axis), or quantified using polynomial contrasts. For transformed data, it must be remembered that the SED applies on the transformed scale and cannot be back-transformed. However, confidence intervals (CIs) for treatment means or differences on the transformed scale can be back-transformed for presentation on the scale of the original data.

At this point in the analysis, multiple comparison tests (e.g. Duncan's multiple range test) are often used to examine all factor level combinations in the table of predicted means. Much has been written on the limitations of the various multiple-comparison tests in the context of entomology (Perry, 1986; Bondari, 1999), plant pathology (Madden, 1982; Gilligan, 1986), weed research (Cousens, 1988) or general agricultural experimentation (Gates, 1991; Pearce, 1993). These papers include many examples where the naive use of a multiple comparison test has obscured the conclusions of an experiment. Common criticisms are that the tests produce contradictory results and lead to unclear conclusions. These authors all agree that multiple comparison tests are inappropriate for analysis of experiments with a factorial treatment structure, and that interpretation of the patterns in the main effects/interactions found to be significant is more informative. In particular, where factors have an underlying numeric structure, it is far better to examine the response across increasing factor levels explicitly using polynomial contrasts, for example, rather than via contiguous paired comparisons. Where the authors disagree is in the analysis of experiments with a completely unstructured set of treatments, for example Pearce (1993) suggests that multiple comparison tests may then sometimes be useful to select an overall "best treatment". However, Perry (1986) strongly argues that, even in this case, multiple comparison tests are less useful than consideration of the structure of the whole set of treatment effects by graphical methods in order to identify groups of similar treatments.

Reporting statistical results in scientific papers

In the same way that the methods section of a scientific paper should allow a fellow scientist to repeat the experiment, the description of the statistical analysis should enable the reader to understand the statistical methods used sufficiently to repeat the analysis. The treatment structure and the experimental design should be stated clearly, as these determine the structure of the analysis. It is also useful to reference the statistical package used. The method of analysis and any diagnostic checks used to verify the assumptions of the analysis should be clearly stated. Significant terms in the treatment model (main effects and interactions) with the corresponding F-statistics and probability levels (e.g. $F_{1,20} = 135.50$, $p<0.001$) should be listed unless the full ANOVA table can be shown. Tables or graphs of appropriate treatment means (i.e. those relating to the simplest model), with SEDs, LSDs or CIs (and the residual degrees of freedom on which they are based), can be used to show the pattern of response to different treatment factors.

Acknowledgements

We thank Bruce Fitt and Julie Steed for permission to use the oilseed rape field experiment data and design details from one of their experiments and Alan Todd for analysis of these data. We also thank Bruce Fitt for comments that improved the text. The work was funded by the Biotechnology and Biological Sciences Research Council.

References

Bailey R A (1984) Quasi-complete Latin squares – construction and randomization. *Journal of the Royal Statistical Society, Series B* **46**: 323-334.

Bondari K (1999) Interactions in entomology: multiple comparisons and statistical interactions in entomological experimentation. *Journal of Entomological Science* **34**: 57-71.

Cochran W G & Cox G M (1957) *Experimental Designs (second edition)*. Wiley, New York.

Cousens R (1988) Misinterpretations of results in weed research through inappropriate use of statistics. *Weed Research* **28**: 281-289.

David O & Kempton R A (1996) Designs for interference. *Biometrics* **52**: 597-606.

Diggle P J, Liang K V & Zeger S L (1994) *Analysis of Longitudinal Data*. Clarendon Press, Oxford.

Gates C E (1991) A users guide to misanalyzing planned experiments. *Hortscience* **26**: 1262-1265.

Gilligan C A (1986) Use and misuse of the analysis of variance in plant pathology. *Advances in Plant Pathology* **5**: 225-261.

Madden L V (1982) Considerations for the use of multiple comparison procedures in phytopathological investigations. *Phytopathology* **72**: 1015-1017.

Mead R (1988) *The Design of Experiments: Statistical Principles for Practical Application*. Cambridge University Press, Cambridge.

Mead R, Curnow R N & Hasted A (1993) *Statistical Methods in Agriculture and Experimental Biology*. Chapman & Hall, London.

Payne R W (2003) *The Guide to GenStat Part 2: Statistics*. VSN International, Hemel Hempstead, UK.

Pearce S C (1993) Data-analysis in agricultural experimentation. 3. Multiple comparisons. *Experimental Agriculture* **29**: 1-8.

Perry J N (1986) Multiple-comparison procedures – a dissenting view. *Journal of Economic Entomology* **79**: 1149-1155.

Piepho H P, Büchse A & Emrich K (2003) A hitchhiker's guide to mixed models for randomized experiments. *Journal of Agronomy and Crop Science* **189**: 310-322.

Snedecor G W & Cochran W G (1989) *Statistical Methods (eighth edition)*. Iowa State University Press, Ames.

Steed J M, Fitt B D L, Gladders P & Sutherland K G (1999) Optimising fungicide timing for control of light leaf spot (*Pyrenopeziza brassicae*) on winter oilseed rape in the UK. *Protection and production of combinable break crops, Aspects of Applied Biology* **56**: 117-122.

Steel R G D & Torrie J H (1990) *Principles and Procedures of Statistics: A Biometrical Approach (third edition)*. McGraw-Hill, New York.

Whitaker D, Williams E R & John J A (2002) *CycDesigN: A package for the Computer Generation of Experimental Designs, Version 2.0*. CSIRO, Canberra.

Sue Welham and *Suzanne Clark* are both statistical consultants working in the Biomathematics and Bioinformatics Division at Rothamsted Research, Harpenden, UK.

Sue Welham has worked at Rothamsted Research as a statistical consultant since 1987, except for a 3-year sabbatical to pursue PhD studies at the London School of Hygiene & Tropical Medicine during 2000-2003. Her research interests are in REML analysis of mixed models, especially smoothing spline models and extensions to spatio-temporal data, and in developing statistical software for these methods. Within Rothamsted, she previously worked mainly with plant pathologists, but now works in all areas requiring analysis of mixed models.

Suzanne Clark has been a statistical consultant, primarily for the entomologists and nematologists at Rothamsted, for 25 years. She advises staff and students on a wide range of statistical issues including planning and design of field, glasshouse, controlled environment and laboratory experiments, data analysis and presentation of results.

Similar articles that appeared in *Outlooks on Pest Management* include – 2006 **17(2)** 88