European Journal of Soil Science

European Journal of Soil Science, January 2018, **69**, 126–139 doi: 10.1111/ejss.12511

Analysis of variance in soil research: let the analysis fit the design

R. Webster[a] & R. M. Lark[b*]

[a]*Rothamsted Research, Harpenden AL5 2JQ, UK, and* [b]*British Geological Survey, Keyworth, Nottingham NG12 5GG, UK*

Summary

Sound design for experiments on soil is based on two fundamental principles: replication and randomization. Replication enables investigators to detect and measure contrasts between treatments against the backdrop of natural variation. Random allocation of experimental treatments to units enables effects to be estimated without bias and hypotheses to be tested. For inferential tests of effects to be valid an analysis of variance (anova) of the experimental data must match exactly the experimental design. Completely randomized designs are usually inefficient. Blocking will usually increase precision, and its role must be recognized as a unique entry in an anova table. Factorial designs enable questions on two or more factors and their interactions to be answered simultaneously, and split-plot designs may enable investigators to combine factors that require disparate amounts of land for each treatment. Each such design has its unique correct anova; no other anova will do. One outcome of an anova is a test of significance. If it turns out to be positive then the investigator may examine the contrasts between treatments to discover which themselves are significant. Those contrasts should have been ones in which the investigator was interested at the outset and which the experiment was designed to test. Post-hoc testing of all possible contrasts is deprecated as unsound, although the procedures may guide an investigator to further experimentation. Examples of the designs with simulated data and programs in GenStat and R for the analyses of variance are provided as File S1.

Highlights

- Replication and randomization are essential for sound experimentation on variable soil.
- Analyses of variance of data from experiments must match the experimental designs.
- Experiments should be designed to answer preplanned questions and test hypotheses.
- Efficiency can be gained by blocking and factorial combinations of treatments.

A little history

In 1843 John Lawes, the then owner of the Rothamsted estate in Hertfordshire, England, and his newly appointed scientist, Henry Gilbert, planned their experiment on Broadbalkfield to test and compare the responses of winter wheat to various combinations of fertilizers. The experimental treatments were applied to long narrow strips of land running the length of the field, which were divided in a perpendicular direction into sections. Lawes and Gilbert weighed the yields, and they sampled both the crop and the soil in every plot in every section so as to measure the off-take of nutrients and the nutrient status of the soil. A few years later they laid down similar experiments on spring barley (Hoosfield, in 1852) and a meadow (Park Grass, in 1856), both of which are still running. They also meticulously recorded the weather. Rothamsted Research (2006) has summarized the history and main findings of these long-term experiments in its guide.

By the end of the First World War, during which Rothamsted began to receive money from the British government for its research, a huge body of data had accrued from these long-term experiments, and in 1919 R.A. Fisher was appointed to analyse the data and make sense of them.

Fisher soon realized that without replication, which was the situation on Park Grass, he could not discover how variable was the response to any one treatment. The treatments on Broadbalk were replicated, but because the different plots for each treatment lay in a

Correspondence: R. Webster. E-mail: richard.webster@rothamsted.ac.uk
*Present address: School of Biosciences, The University of Nottingham, Sutton Bonnington Campus, Sutton Bonnington, Leicestershire LE12 5RD, UK.

This paper was invited for the 70th Anniversary of the British Society of Soil Science and was received on 24 July 2017; revised version accepted 7 November 2017

single strip he could not separate the effects of the treatments from the soil's natural variation as expressed in differences between the strips. This natural variation and the treatment effects are said to be confounded. The treatments on the spring barley experiment were replicated on plots that were separated from one another but in a way that might be confounded with the natural variation in the field. So, again, it was not possible to estimate the effects of the fertilizers alone.

Having recognized the serious shortcomings of those old trials, Fisher formalized and systematized what had, hitherto, been inconsistently and erratically applied elements of experimental design. One was replication, present in some of the experiments but not all, and necessary to provide information on the variation in responses. The other was randomization, necessary to avoid the bias that could arise if treatment effects are confounded with sources of variation that are uncontrolled and might be unknown. Fisher devised the analysis of variance (ANOVA) to separate the sources of variation in data from such experiments, to estimate quantitatively the effects of different treatments and to provide inferential tests to judge whether the observed differences could have arisen by chance rather than as results of the imposed treatments. Fisher also introduced blocking to remove effects such as trends across experiments. Trends of this kind do not introduce bias if the experimental design is randomized, but blocking improves the sensitivity of the experiment to detect treatment effects against the background variation represented by the trends.

Fisher's principles of experimental design and the concomitant analysis of variance are as valid today as they were 90 years ago. They have been the foundation of agronomic practice ever since, and statisticians collaborate with agronomists to ensure that designs will produce data that can be analysed to answer the questions put at the outset. Numerous text books are available to guide practitioners; two that we can recommend unreservedly are that by Snedecor & Cochran (1989) and the more recent book by Mead *et al.* (2003). Cochran & Cox (1957) remains a standard text. You might like also to see the Statistical Checklists prepared by Jeffers (1978).

Sadly, many of today's soil scientists are working without the guidance or collaboration of statisticians. One consequence is that they often plan experiments and surveys that cannot or are unlikely to answer their questions; or having designed the experiments soundly they vitiate the potential of the experiments to answer the questions by improper sampling. Or they see opportunities to answer new questions that were not envisaged when the original experiments were planned, either by themselves or by other scientists, yet fail to appreciate the limitations inherent in the designs. A further consequence is that despite having designed their experiments and surveys well they analyse the data from them incorrectly. All too often they load their data into a statistical package, press a few buttons on a menu without understanding, and copy the output into their scripts.

We write in this critical vein from our experiences as advisors to the journal's editors in the last few years, and from the experience of the journal's statistical advisory panel. It is no exaggeration to state that most of the papers on which the editors have sought advice have embodied one or more of the above failings. In the first set of circumstances we have felt obliged to judge the results of little worth and to advise the editors to reject the papers. To paraphrase one of R.A. Fisher's remarks, it has been like conducting post-mortems, only to say what the experiments died of. In some instances we have asked for further sampling. In the second set we have seen that redemption is often possible by fresh and correct analysis of the data.

In one short article we cannot describe all that investigators should do. Instead, we focus on the specific matter, namely analyses of variance that follow from the designs, and in particular on the most frequent mismatches between design and analysis. At the best such mismatches lead to loss of information and so to waste of the effort required to do the experiment. At worst, the inferences made from the analysis are unsafe and lead to bad decisions. We have already remarked on this in an editorial (Webster *et al.*, 2016). In the comic opera *The Mikado* by W.S. Gilbert and Arthur Sullivan the Mikado himself demands that the punishment fit the crime. Here we demand that the analysis fit the design.

Designs

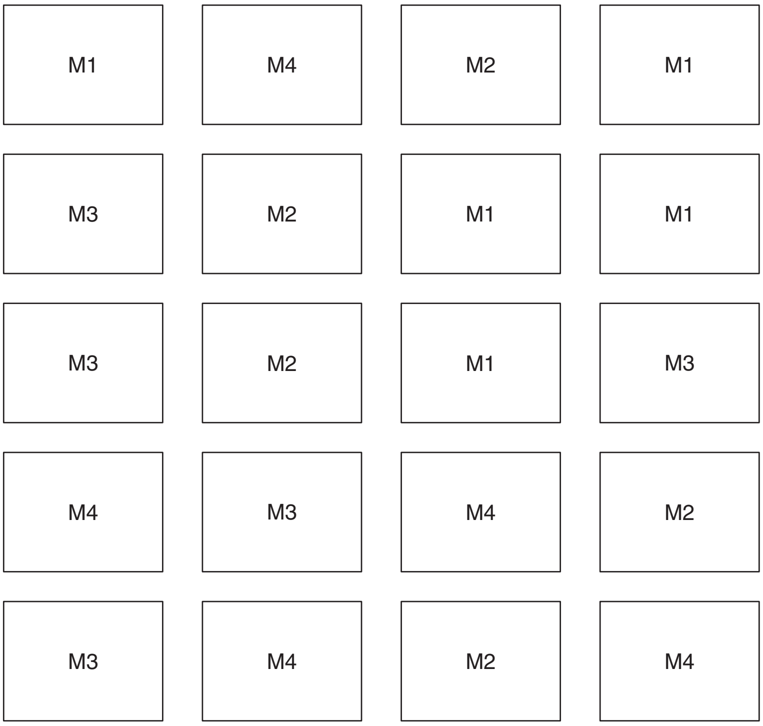
We describe in detail below the commonest and most straightforward designs, starting with the simplest, completely randomized schemes, introducing blocking, and progressing to factorial and then split-plot designs. We have provided examples of these designs with simulated data, together with programs in GenStat and R for the correct analyses of variance and the output from those analyses in the File S1.

Completely randomized (CR) design

We begin with the simplest design. Suppose that investigators wish to compare the effects of several manure treatments on some property of the soil, say the microbial biomass, which we shall denote z. They replicate their treatments and assign them to the experimental plots in a completely randomized and independent way. Let there be n_1 treatments, each replicated n_2 times, so that there are $N = n_1 \times n_2$ plots, or units, of the design. Treatments are allocated to plots independently and at random. This means that the probability that the first plot in the experiment is allocated to the jth treatment is n_2/N, equivalently $1/n_1$. Subsequently when n_j replicates of the jth treatment remain to be assigned, the probability that any one of the N_u plots that have still to be assigned a treatment will ultimately receive treatment j is n_j/N_u. Figure 1 shows one outcome of such assignment in which $n_1 = 4$ and $n_2 = 5$.

The files exp1.* in the File S1 contain data with this design and the programs for analysing them.

The analysis of variance for this design appears in Table 1. Note that this presentation of the analysis of variance, and that for subsequent designs, holds for the balanced case in which the numbers of replicates of the treatments are equal. The texts to which we have referred provide further information on analysis in the unbalanced case, but the topic is beyond the scope of the paper. The total mean



M1	M4	M2	M1
M3	M2	M1	M1
M3	M2	M1	M3
M4	M3	M4	M2
M3	M4	M2	M4

Figure 1 An example layout of a completely randomized balanced experimental design in which five replicates of each of four manure treatments, M1, M2, M3 and M4, are independently and randomly allocated to plots.

Table 1 Analysis of variance for n_1 treatments replicated n_2 times in a completely randomized (CR) design

Source	Degrees of freedom	Mean squares	Parameters estimated	*F* ratio
Between treatments	n_1-1	B	$\sigma^2_{\mathrm{W}}+n_2\sigma^2_{\mathrm{B}}$	B/W
Within treatments (residual)	$n_1(n_2-1)$	W	σ^2_{W}	
Total	n_1n_2-1	T		

square is T:

$$T=\frac{1}{n_1n_2-1}\sum_{j=1}^{n_1}\sum_{i=1}^{n_2}\left(z_{i,j}-\overline{z}\right)^2, \tag{1}$$

where $z_{i,\,j}$ is the measured response of the *i*th replicate of the *j*th treatment and $\overline{z}$ is the mean response over all n_1n_2 plots. One can see that this quantity is a variance, the variance of the plot responses. The divisor of the sum of squares, n_1n_2-1, is called the degrees of freedom in Table 1. It can be regarded as the number of independent pieces of information about the variation of the plot responses provided by the data. There are n_1n_2-1 degrees of freedom rather than n_1n_2 because each plot response is compared wiith the overall mean estimated from all the data. Because

$$\sum_{j=1}^{n_1}\sum_{i=1}^{n_2}\left(z_{i,j}-\overline{z}\right)=0,$$

it follows that, when we know the values of n_1n_2-1 differences in the summation, the last one is fixed and so provides no new information.

The within-treatment mean square, W, is computed as

$$W=\frac{1}{n_1\left(n_2-1\right)}\sum_{j=1}^{n_1}\sum_{i=1}^{n_2}\left(z_{i,j}-\overline{z_j}\right)^2, \tag{2}$$

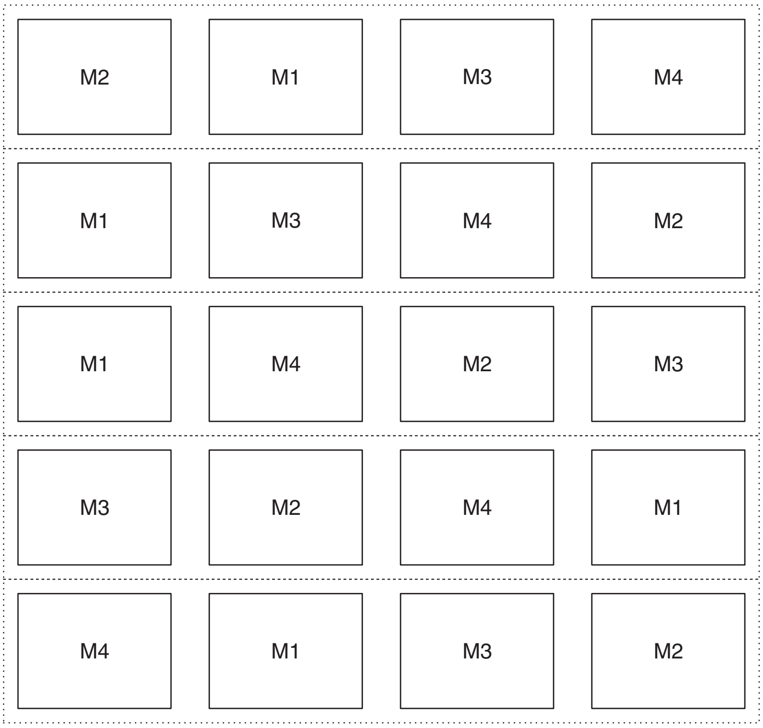
where$\overline{z_j}$ is the average response of all plots in the *j*th treatment. The value estimated by W is the variance of plot responses within the treatments (i.e. the variance about the treatment means). This quantity is σ^2_{W} in Table 1. It has $n_1(n_2-1)$ degrees of freedom in this simple balanced case because each of the n_1 treatments contributes n_2-1 degrees of freedom from the independent variations about the mean of its n_2 replicates, from which the treatment mean is estimated.

The between-treatment mean square, called B in Table 1, is computed for this simple balanced case as

$$B=\frac{1}{n_1-1}\sum_{j=1}^{n_1}n_2\left(\overline{z_j}-\overline{z}\right)^2 \quad . \tag{3}$$

This is equivalent to the sum, over all plots, of the squared difference between the corresponding treatment mean and the overall mean, divided by the number of independent variations among the treatment means.

The residual mean square in an analysis of variance is a direct estimate of a variance component. In general, however, mean squares estimate combinations of more than one variance component. Table 1 shows that B estimates $\sigma^2_{\mathrm{W}}+n_2\sigma^2_{\mathrm{B}}$. The quantity σ^2_{B} is the variance among the treatment means. If there were no differences between the treatments then this quantity would be zero, and, as can be seen in the table, B and W would both estimate σ^2_{W}, and



M2	M1	M3	M4
M1	M3	M4	M2
M1	M4	M2	M3
M3	M2	M4	M1
M4	M1	M3	M2

Figure 2 An example layout of a randomized blocked experimental design in which the plots are grouped in blocks of four (separated by the dotted lines) and one replicate of each of four manure treatments, M1, M2, M3 and M4, is independently and randomly allocated to a plot within each block. There are five blocks in total, separated by dotted lines in the Figure.

the ratio $F = B/W$ in the table would have an expected value of 1. We use the standard notation of the Roman letter s for an estimate of the underlying quantity σ, so by s^2_{W} we denote the estimate of σ^2_{W} provided by W in Table 1.

Apart from separating the sources of variation in the experiment and providing quantitative values of the variances attributed to those sources, the analysis enables us to draw inferences. If the responses in z to the treatments differ from one another then we should expect the ratio B/W to exceed 1. But B/W could exceed 1 purely through random variation; so how can we tell that we have a real effect of the treatments? We do so by putting forward the 'null hypothesis', often designated H0 in statistics textbooks. It is the hypothesis that there are no differences, and we consider the strength of evidence against it. That evidence is the magnitude of B/W in relation to the distribution of F if the null hypothesis were true. We can tell we have a real effect because, as a result of our design, B and W would be independent estimates of σ^2_{W} if the null hypothesis were true. It follows from the independent random allocation of treatments to plots, and it appears in the ANOVA table in the way that the $n_1 n_2 - 1$ total degrees of freedom are partitioned into the between-treatment and within-treatment (residual) degrees of freedom.

In these circumstances the variance ratio has the F distribution under the null hypothesis and the shape of the distribution that depends on the degrees of freedom for the numerator and denominator of the ratio. One can therefore compute the probability that an F ratio as large or larger than the value observed in the table would arise under the null hypothesis through random variation. The smaller is this probability, or P-value, the stronger is the experimental evidence that we should reject the null hypothesis and say that the treatments have produced different responses. It is now a short step to the common notion of statistical significance. It is conventional to take $P = 0.05$ as a threshold. If P exceeds 0.05, investigators accept the null hypothesis. Otherwise, with $P \leq 0.05$ they declare that the observed differences are 'significant', and they decorate their tables of means with stars, which again we deprecate! One may choose some other value of P depending largely on how serious it would be to come to a false conclusion.

Inference from the analysis of an experiment like that above is based on assumptions about the distribution of random quantities under the null hypothesis that are justified by that design, the way it was laid out in the field, glasshouse or laboratory, and on the numbers of the degrees of freedom for the variance ratio. In this sense the analysis (and ANOVA table) matches the design.

Randomized complete block (RCB) design

Where investigators know of or suspect trends in fertility, drainage or pollutants that might affect their results they typically replicate their treatments in blocks. In the simplest case each treatment is replicated once and only once in each block. The allocation of treatments within the blocks is carried out independently and at random. Figure 2 shows one realization of an RCB design for four treatments and five blocks, and so the same total number of replicates as the completely randomized case in Figure 1. The blocks are separated by the dotted lines; notice that in each block there is one plot for each of the n_1 treatments. The blocks in this figure are laid out as rows across the experimental layout and so would be suitable if a trend in soil properties was known or suspected to occur from the top to the bottom of the site.

Table 2 Analysis of variance for n_1 treatments replicated n_2 times in a randomized complete block (RCB) design

Source	Degrees of freedom	Mean squares	Parameters estimated	*F* ratio
Blocks	n_2-1	*A*	$\sigma^2_{\mathrm{W}} + n_2\sigma^2_{\mathrm{A}}$	*A*/*W*
Between treatments	n_1-1	*B*	$\sigma^2_{\mathrm{W}} + n_2\sigma^2_{\mathrm{B}}$	*B*/*W*
Within treatments (residual)	$(n_1-1)\times(n_2-1)$	*W*	σ^2_{W}	
Total	n_1n_2-1	*T*		

The files exp2.* in the File S1 contain data with this design and the programs for analysing them.

The analysis of variance for this design, still with n_1 treatments each replicated once in each of n_2 blocks, appears in Table 2. Here σ^2_{W} and σ^2_{B} are the underlying variances for plots and treatments as before. There is an additional line in the table for the between-block mean square with n_2-1 degrees of freedom; σ^2_{A} is the variance between blocks. The total degrees of freedom and the treatment degrees of freedom are unchanged from Table 1, but there are n_2-1 fewer residual degrees of freedom. This follows from simple arithmetic, but it also indicates that the random allocation of treatments to plots is more constrained in the RCB design than in the CR design (once one plot in block *k* has been assigned to the *j*th treatment we know that no other plot in the block will receive it). For this reason there is somewhat less information in the residual mean square than in the CR design with the same number of plots and treatments.

Where does the between-block variance come from? It is natural variation in the experimental environment that appears as between-block rather than within-block variation. If blocking were not undertaken then this variation would be part of the residual variance, σ^2_{W}. This means that, if the between-block variance is large, then we reduce the residual variance and so should increase the variance ratio *B/W*, making the experiment and analysis more sensitive for comparing the differences between the treatments. This is why blocking, appropriately planned, should be advantageous. Snedecor & Cochran (1989) provide formulae for calculating the efficiency of blocking. At its simplest they calculate it as the ratio of the residual variances:

$$\text{Efficiency} = s^2_{\mathrm{CR}}/s^2_{\mathrm{RB}} \ , \qquad (4)$$

wheres^2_{CR} is the residual variance on the assumption that the design was completely randomized (CR), whereas s^2_{RB} is the residual variance of the RCB design. You can find further details of the calculation on pages 263 and 264 of Snedecor & Cochran (1989).

An efficient blocking design is evidently one in which the differences between the blocks are larger than the variation within the blocks. In practice one might achieve this by keeping the blocks compact, although in a field where there is a strong trend in the soil or environment in one direction rectangular blocks with the long side perpendicular to the direction of the trend would be preferred. It is important to pay attention to the structure of the blocks, because, as above, there is a small penalty for blocking from the reduced residual degrees of freedom, and this will be worth paying only if there are real differences between the blocks.

The variance ratio *A/W* appears in Table 2, and one could use it to test the null hypothesis that the between-block variance, σ^2_{A}, is zero. That would be of interest only in that it shows whether the blocking is better than random assignment of plots to blocks. Sometimes, however, the scientist, having found that the evidence for a difference among the blocks is weak, ignores the blocking and reports an analysis of variance appropriate for a CR design. Such an analysis does not fit the design. The scientist might try to justify that analysis because the blocks have been shown not to differ, but that misses the point. What the correct analysis shows us, and shows explicitly in the ANOVA table, is how the actual allocation of treatments to plots was undertaken; it shows that in the RCB case we have $(n_1-1)\times(n_2-1)$ degrees of freedom, not $n_1(n_2-1)$. In short, the correct analysis reports the reduction, albeit small, in information about the residual variance that follows from the constraints of blocking. The extra n_2 residual degrees of freedom in the analysis as if the design were completely randomized means that, other things being equal, a given variance ratio appears to offer stronger evidence against the null hypothesis. This inference would be unsafe, however, because the quoted degrees of freedom would not describe the actual randomization. In practice this would mean that the variance ratio for a treatment effect would be compared with the wrong distribution of the *F* statistic. The analysis would not fit the design.

The Austrian philosopher Ludwig Wittgenstein was once impressed by an account of a trial that took place following a car accident in Paris. During the trial, models were used to represent the positions of the vehicles involved at the time of the collision (Kenny, 2005). Inspired by this, he developed his picture theory, by which a logical proposition is equivalent to a picture of a state of affairs in the world. Such a proposition may take different forms. It may, for example, be spoken, written or drawn. Let us apply the idea in the present context to the design of field experiments.

Consider an experiment that has been performed according to an RCB design. The design could be illustrated with a diagram such as Figure 2. More often in scientific papers the designs are described in words in *Methods* sections. The equivalent to Figure 2 would be ‘The n_1 treatments were allocated independently and at random within each of n_2 blocks’. Our contention is that the correct analysis of variance table for the experiment, as shown in Table 2, is one more way in which we may express the same proposition. The partition of the sum of squares between rows of the table represents the sources of variation that the experimental design uniquely induces, and the numbers of degrees of freedom show how many blocks and replicates were used as surely as does Figure 2 or the verbal statement.

That is one reason why this journal asks its authors to provide full ANOVA tables. The request is sometimes misinterpreted as a request for a table of only a set of variance ratios and corresponding *P*-values, but that is not what is required. The journal requires a table like Tables 1 or 2 shown here, because such a table represents the design definitively. When assessing an experiment both the reviewers and, ultimately, readers must be able to see that the experiment as described in the methods section accords with the ANOVA reported in the results.

Factorial designs

When an investigator is interested in the effects of several factors it is much more efficient to include them in a single experiment than in a series of separate experiments, one for each factor. This was recognized by Fisher (1926) who wrote:

> *No aphorism is more frequently repeated in connection with field trials, than that we must ask Nature few questions, or, ideally, one question, at a time. The writer is convinced that this view is wholly mistaken. Nature, he suggests, will best respond to a logical and carefully thought out questionnaire; indeed, if we ask her a single question, she will often refuse to answer until some other topic has been discussed*.

Yates (1937) set out the principles of factorial designs in his *Technical Communication 35*, which became the guiding text for fertilizer trials for many years. More recently, Carmer & Walker (1982) have urged investigators to take this course.

To illustrate the principles of the design and corresponding analysis we take a simple example with three factors, the major plant nutrients, nitrogen (N), phosphorus (P) and potassium (K). Factors are each applied at two or more 'levels'; in this example we assume that the nutrient is either applied or not (two levels). There are therefore $2^3 = 8$ combinations of factor levels; these are our treatments. The treatments must be replicated between units (plots in this case) according to a suitable design, and analysed in accordance with that design. One might use CR or RCB designs, as in the examples already discussed.

Let us assume that there are, as before, n_2 replicates arranged in a CR design. We could analyse the data as set out in Table 1 with $8-1=7$ degrees of freedom for the treatments. This analysis would be quite correct, but it would not be very informative. If we found that the treatments were significantly different then how should we interpret this finding in terms of all our three factors? The factorial design allows us to do this. We can partition the sum of squares due to differences among the treatments into what are called main effects and interactions. There are three main effects in our example, the differences between treatments with contrasting levels of N is one such, and the other main effects are due to P and K. If these effects simply add to one another then all of the treatment sum of squares will be accounted for by the sums of squares for the three main effects. If, in contrast, the difference between plots that receive N and those that receive none is not the same on plots that receive K and those that receive no K then the factors K and N are said to interact. One can see that there are three such interactions in our example: N·P, N·K and P·K. To complicate matters further, if the N·K interaction differs between plots that receive P and those that receive none, then there is a three-way interaction N • K • P. Note that we could express the same three-way interaction in terms of an effect of, for example, the level of N on the P·K interactions, so there is just one three-way interaction in a factorial experiment with three factors. We use this 'dot' convention to indicate interactions as established by Wilkinson & Rogers (1973).

Table 3 Three-way analysis of variance for three factors, N, P and K, each at two levels replicated n_2 times in a completely randomized (CR) design

Source	Degrees of freedom	Parameters estimated by mean squares	*F* ratio
Between treatments	7	$\sigma_W^2 + n_2\sigma_B^2$	
N	1	$\sigma_W^2 + n_2\sigma_N^2$	
P	1	$\sigma_W^2 + n_2\sigma_P^2$	
K	1	$\sigma_W^2 + n_2\sigma_K^2$	
N • P	1	$\sigma_W^2 + n_2\sigma_{NP}^2$	
N • K	1	$\sigma_W^2 + n_2\sigma_{NK}^2$	
P • K	1	$\sigma_W^2 + n_2\sigma_{PK}^2$	
N • P • K	1	$\sigma_W^2 + n_2\sigma_{NPK}^2$	
Within treatments (residual)	$8\times(n_2-1)$	σ_W^2	
Total	$8\times n_2-1$	σ_T^2	

Table 3 sets out the ANOVA for our example. Note that each main effect has a single degree of freedom; this is because there are two levels of each factor, and so the main effect consists of just the difference between the responses to these levels. In general, a factor with U_1 levels has U_1-1 degrees of freedom for its main effect. Similarly, the two-way interactions each have one degree of freedom; in general, two factors with U_1 and U_2 levels have an interaction with $(U_1-1)\times(U_2-1)$ degrees of freedom. Equally the three-way interaction has 1 degree of freedom in our example. In the general case where the third factor has U_3 levels, the three-way interaction has $(U_1-1)\times(U_2-1)\times(U_3-1)$ degrees of freedom. The reader will note that in our example the sum of the degrees of freedom for the main effects and interactions is 7, the same as the treatment degrees of freedom. The treatment degrees of freedom are partitioned between main effects and interactions, as is the treatment sum of squares.

The quantity σ_W^2 in Table 3 is the underlying variance among the plots receiving the same combination of treatments, and σ_N^2, σ_P^2, ..., σ_{NPK}^2 are the variances attributed to the nutrients and their combinations. The *F* ratio for any one entry is:

$$F = \frac{\text{mean square for the treatments}}{\text{residual mean square}} \;. \tag{5}$$

The standard error of any of the treatment means is:

$$\mathrm{SE}_{\text{treatment}} = \sqrt{\text{residual mean square}/n_2} \;. \tag{6}$$

Where the investigator goes from there depends very much on the outcome of the analysis. If it turns out that the interactions, especially the threefold interaction of N, P and K, are non-significant and only the main effects of the three nutrients are significant, the

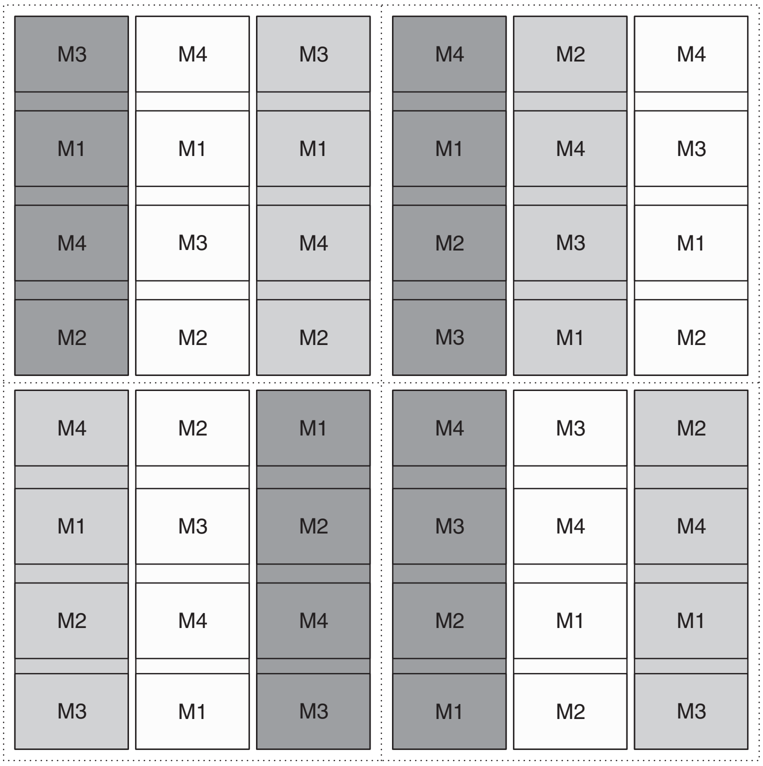


Figure 3 An example layout of a split plot design with blocks. Three main plots are in each block, and one replicate of each of three levels of an irrigation factor, I1, I2 and I3, is independently and randomly allocated to a main plot within each block. The three levels of the irrigation factor are distinguished in this figure by dark grey, light grey or white shading. Within each main plot are four subplots and one replicate of each of four manure treatments, M1, M2, M3 and M4, is independently and randomly allocated to a subplot within each main plot.

investigator may choose to focus on the main effects, i.e. on the means of plots receiving each of the N, P and K averaged over all combinations that include them. Their standard error is

$$\mathrm{SE}_{\text{main effect}} = \sqrt{\text{residual mean square}/4n_2}\ . \qquad (7)$$

The quantity 4 appears in the denominator because, in the example, n_2 replicates of four treatments contribute to the estimate of the mean response for each level of one of the factors.

We cannot consider here all the possible outcomes and their consequences; rather we must leave readers to pursue them elsewhere. Again, we recommend Snedecor & Cochran (1989).

We include this account of factorial designs and analysis because all too often in papers submitted to the journal the analysis does not match the design. Some authors, having undertaken an experiment according to a factorial design, proceed to analyse it in a series of one-way analyses for each of the main effects. This is bad practice for two reasons. If all the data from the experiment are analysed in this way then the influence of those main effects not considered in a particular analysis will inflate its residual mean square. Further, when there is a substantial interaction between factors the main effect may be small or negligible, even though the factor is an important one. This is our interpretation of what Fisher means by saying that nature ‘may refuse to answer’ a particular question ‘until some other topic has been discussed’. If the design is factorial then the analysis should be so as well, otherwise it is very likely that substantial information will be lost.

Split plots

Split-plot designs are common in agricultural experimentation. There are two general circumstances in which they are used. The first is a factorial experiment in which one of the factors can be replicated only between fairly large plots for logistical reasons. A typical example is where one of the factors is an irrigation or drainage treatment. Large plots are needed for these, but it would not be feasible to replicate such plots in factorial combination with several fertilizer treatments as above. The experiment would require too large an area to manage. The solution is to replicate the irrigation factor between appropriate large plots (main plots in the jargon), and then to divide each main plot into subplots, one subplot for each level or combination of levels of the remaining factors, which are allocated to subplots at random.

Let us suppose that the four manure treatments of Figure 1 (M1, M2, M3, M4) are to be combined in an experiment in which there are three irrigation treatments (I1, I2, I3), say no irrigation, irrigation when the soil has dried to half its available water capacity and irrigation at regular intervals regardless of the water deficit. Figure 3 shows a possible layout on the ground with the irrigation treatment replicated between main plots in the blocks, and the manure treatments replicated between subplots within each main plot.

How would the data from this experiment be analysed? There are 12 treatments (combinations of the four levels of the manure factor and the three levels of the irrigation factor). The treatments are replicated in four blocks. One might think that Table 4 would partition the degrees of freedom for the ANOVA; the design is after

Table 4 Incorrect partial analysis of variance table for the factorial experiment with manure and irrigation factors illustrated in Figure 3

Source	Degrees of freedom
Between blocks	3
Between treatments	11
Manure	3
Irrigation	2
Manure • irrigation	6
Residual	33
Total	47

Table 5 Analysis of variance for the split plot experiment with three levels of the irrigation factor replicated between main plots within blocks, and four levels of the manure factor replicated between subplots within each main plot

Source	Degrees of freedom	Mean squares	*F* ratio
Main plots			
Block	3	B_{B}	$B_{\mathrm{B}}/W_{\mathrm{MP}}$
Irrigation	2	B_{I}	$B_{\mathrm{I}}/W_{\mathrm{MP}}$
Main plot error	6	W_{MP}	
Subplots			
Manures	3	B_{M}	$B_{\mathrm{M}}/W_{\mathrm{SP}}$
Irrigation • manures	6	B_{IM}	$B_{\mathrm{IM}}/W_{\mathrm{SP}}$
Subplot error	27	W_{SP}	
Total	47	T	

The subscripts are B for block, I for irrigation, M for manures, MP for main plot, SP for subplot, and MPE and SPE denote the main-plot and subplot errors.

all a factorial one. However, an analysis with that structure would be wrong; the table does not match the design. To see this, reflect on the basic units of the experiments, the subplots; there are 12 of them in each block. The ANOVA structure in Table 4 implies that there are no constraints on the randomization of the 12 treatments between subplots within each block, but that is not the case. If we are told that a plot in the top left corner of a block has treatment I3-M4 we can know, first, that all plots in the same main plot receive level I3 of the irrigation factor, and second, that no other subplot in the main plot receives level M4 of the manure treatment. In short, Table 4 fails to show that the levels of the irrigation factor were allocated to the main plots while the levels of the manure factor were allocated to subplots within the main plots.

Table 5 sets out the correct analysis for this experiment with the three levels of the irrigation factor randomly allocated between main plots in each of four blocks, and the four levels of the manure factor randomly allocated to the subplots within each main plot.

The files exp3.* in the File S1 contain data with this design and the programs for analysing them.

Notice how the *F* ratios are calculated in Table 5. The denominator for the irrigation *F* ratio is the main-plot error mean square. That for the manures and the interaction between the irrigation and manures is the subplot error mean square. In such a design the subplot error variance is smaller than the main-plot error variance. These variances follow through to different standard errors for the means. In this example the manure treatments are compared more sensitively than the irrigation treatments. If the data from this experiment were mistakenly analysed as in Table 4 then one would underestimate the main-plot error variance and overestimate the subplot variance.

In an experiment like the one above, the treatments, say manure and irrigation, are laid out in split-plot designs from the start. Although such experiments are not always correctly analysed in papers submitted to the journal, problems more often arise when split-plots are introduced into experiments later on. Consider an original RCB experiment with four treatments like that above. Let us suppose that the treatments are four different kinds of manure and that the investigator planned to compare rates of respiration in the soil between these treatments. Having seen the results, he or she then introduces a second factor, the soil water potential. Two soil cores are taken from each plot of the original experiment and equilibrated at one of two soil water potentials, and then the respiration rate of each is measured. The plots in such an experiment are not physically split, and authors are sometimes puzzled when we tell them that they have split-plot designs. They need to recognize that in such a situation the experiment has a split-plot design with manures replicated between main plots and the cores extracted from each main plot serve as subplots between which the levels of the water-potential factor are randomized. This should be reflected in an ANOVA table like Table 5. Too often we receive papers in which such experiments are analysed as if they had simple RCB factorial designs.

Sampling within experimental plots

One can rarely measure soil properties of whole plots; almost always the most one can do is to sample the soil and measure the properties of interest on the samples. If one were to take one sample, whether as a single core or a bulked sample from several cores, one would analyse the measurements as above according to the design (i.e. completely randomized or blocked).

However, one might well measure the property on each of several cores from each plot. This would provide information on the variation within the plots, and one could elaborate the analysis of variance accordingly. Suppose that one takes n_3 cores of soil from each and every plot, as illustrated in Figure 4 in which there are $n_1 = 4$ treatments replicated $n_2 = 5$ times in a completely randomized arrangement, and $n_3 = 3$ cores per plot. The correct analysis of variance for this design is set out in Table 6. The quantities σ^2_{W} and σ^2_{B} are the underlying variances between plots within treatments and between treatment means, respectively, and σ^2_{C} is the variance among cores within plots. This table is comparable to one for a split-plot design with cores as the subplots. The difference is that no factor is replicated randomly at the core level. The replication is simply to improve estimates of the plot means. Nonetheless, the between-treatment mean square must be compared with the correct

Table 6 Analysis of variance for n_1 treatments replicated n_2 times on plots in a complete randomized block design with n_3 measurements per plot

Source	Degrees of freedom	Mean squares	Parameters estimated	*F* ratio
Between treatments	n_1-1	B	$\sigma_C^2+n_3\sigma_W^2+n_2n_3\sigma_B^2$	B/W
Between plots within treatments	$n_1(n_2-1)$	W	$\sigma_C^2+n_3\sigma_W^2$	
Between cores within plots	$n_1n_2(n_3-1)$	C	σ_C^2	
Total	$n_1n_2n_3-1$	T		

residual, the between-plots within-treatments mean square, because the treatments are randomized at the plot level.

The standard error of a plot mean is $\mathrm{SE}_{\mathrm{plot}} = \sqrt{C/n_3}$, where C is the variance between cores within plots. If we denote the estimated variance between plots within treatments by s_{W}^2 we obtain the standard error per treatment mean as

$$\mathrm{SE}_{\mathrm{treatment}} = \sqrt{\frac{C}{n_3n_2}+\frac{s_{\mathrm{W}}^2}{n_2}}\ . \tag{8}$$

If the replicates were arranged in blocks then there would be a corresponding additional entry for blocks in the analysis.

Pseudo replication

In the previous example, with the ANOVA as in Table 6, the experimenter recognizes that treatments are replicated and randomized at the plot level, even though measurements are made on n_3 cores in each plot. If, incorrectly, the experimenter treated this design as one with $n_3 \times n_2$ independent replicates of each treatment, it would be a case of what statisticians call ‘pseudo replication’. We introduce the topic of pseudo replication here because many authors of the papers we see commit it either inadvertently or knowingly without appreciating its inferential consequences. We distinguish three situations.

1. The investigator misguidedly regards all $n_2 \times n_3$ observations on each treatment as the units of the design and for a CR design analyses the data as in Table 1. He or she then tests the treatment mean against a residual mean square with $n_1 \times n_2 \times n_3 - n_1$ degrees of freedom. This comprises a form of pseudo replication because the replicates within plots are not true replicates of the experimental treatments. Fortunately, no serious damage is done; once alerted to the mistake the investigator can re-analyse the data correctly according to Table 6.
2. A similar situation arises when a scientist takes either a single core from each plot or bulks multiple cores from each and then splits them into several subsamples for measurement in the laboratory. These replicate measurements cannot be regarded as independent units in the design. They are pseudo replicates. They may be averaged and analysed as in Table 1, or they may be analysed as individual values as in Table 6. In the latter case the variance σ_{C}^2 represents the variance due to subsampling of a single core or composite sample, rather than within-plot variance.
3. Most serious of all is when an investigator takes multiple cores of soil from an experiment that itself has few replicates, perhaps only one, and believes that treating the numerous cores as units will compensate for lack of replication of the main plots and analyses the data according to Table 1. The correct analysis is that exemplified in Table 6. With few true replicates of the treatments, however, the experiment is unlikely to be sufficiently sensitive to reveal any but the biggest and most obvious differences. Here the shortcoming is in the design; the experiment should have been planned with more replication in the field and more resources allocated to its execution.
   The situation arises more often in surveys where investigators want to know how the soil differs from one cultural practice or environment to another. The main difficulty here is in finding sufficient replicates of each kind of practice or environment, especially if access to and travel between them are time-consuming and expensive. What usually happens is that the investigator replicates observations at the few sites that can be reached, often only one of each kind.
   Mean values for the sites actually sampled might be estimated precisely, but differences between practices or environments would not be. If the latter are not replicated, perhaps because replication was impossible, then the investigator can say at the end only by how much the sites themselves differ from one another; any inference about the populations they represent cannot be based on the statistics.

Repeated measurements

The last couple of decades have seen increasing interest in the behaviour of soil over time. Soil scientists have monitored the soil and planned experiments with installations such as static chambers in which to collect gaseous emissions (see, for example, González-Méndez *et al.* (2015) and their repeated measurements of the associated redox potentials from electrodes buried in the soil (González-Méndez *et al.*, 2017)), lysimeters in which to monitor leachates passing through the soil, laboratory reactors in which organic matter is mineralized (e.g. Coban *et al.*, 2016) and microcosms in which to measure the responses of bacteria to imposed treatments over time. The scientists quite properly design their experiments by assigning their treatments to the units, whether chambers, electrodes, lysimeters, reactors or microcosms, with replication and randomization. Then at intervals they make their measurements on every unit. This is especially easy when the measurement is non-invasive, for example by spectrometers. It is also feasible to do so by repeated subsampling of soil from microcosms or field plots. (The soil in long-term experimental plots at Rothamsted has been sampled at intervals over the years since they were first established.)

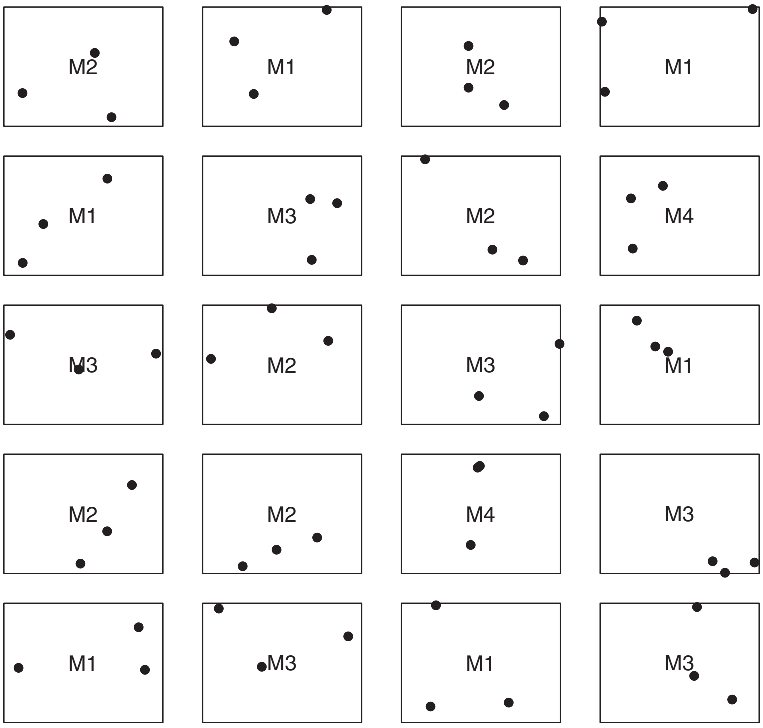


Figure 4 An example layout of the same completely randomized balanced experimental design exemplified in Figure 1 with sites for collection of three soil cores (black discs) independently and randomly located within each plot.

If measurements are made on only two occasions then an appropriate analysis of the data depends on the specific objectives of the experiment. If the variable of interest is the difference between the two observations (e.g. the change in a soil property between the start of a growing season and the end) then the difference may be computed directly for each experimental unit and, being replicated at the level of these units, may be analysed in a straightforward way. If the two observations on each unit are to be analysed together then we have a split-plot design with the chambers, electrodes, lysimeters or microcosms as replicated main plots and the two occasions as subplots within the main plots. One can analyse the data quite correctly as set out in Table 5.

In situations when observations are repeated on the same units, and they are made on more than two occasions, one must take into account possible correlations between the repeated measurements on any one unit. These correlations might depend on the time interval between the observations, which the simple split-plot analysis cannot accommodate. The successive measurements on any one installation cannot be regarded as independent. For the purpose of the statistical analysis the chambers, electrodes, lysimeters or microcosms are the units. The data comprise repeated measurements on those units, and special techniques that take into account the possible correlations are required to analyse them. The techniques often go under the name of 'longitudinal analysis'.

There is no single correct way of analysing repeated measurements, and we cannot delve into the detail of any of them. Webster & Payne (2002), in this journal, reviewed several options. They described in detail one in which the order of correlations was estimated first by an antedependence analysis, as devised by Kenward (1987), and the results of this were then incorporated into an analysis of differences between treatments by residual maximum likelihood (REML). Other options in which the variations in time are modelled as autoregressive processes are available, see again Coban *et al.* (2016).

In whatever way data of repeated measurements are analysed, that way must honour the design. If you wish to investigate processes in the soil over time with fixed installations such as static chambers or lysimeters or in the laboratory with microcosms, then plan your experiments in consultation with a professional statistician and know in advance how you will analyse the data. Of course, you should always know how you will analyse data from any experiment you plan, and for the more straightforward cases you can find recipes in textbooks.

Inferences and comparisons

Orthogonal contrasts

Obtaining a statistically significant result from an ANOVA, say one for which $P < 0.05$, is never the end of an investigation. On its own it is of limited interest. Far more important are the differences between the means: which of the differences contributed to the result? And are they the ones about which the investigator wanted to know when the experiment was designed?

Consider an experiment in which a scientist wants to compare the effects of organic additions to the soil on the respiration rate. The materials to be added are barley straw, wheat straw, cattle slurry and pig slurry. In addition to these four treatments there is a fifth treatment, a control where nothing is added. When this experiment is complete the ANOVA table will include a treatment mean square with four degrees of freedom. This mean square may be compared with the residual mean square to test the null hypothesis that there are no differences in response to the different treatments. Let us suppose that the P-value is so small that the null hypothesis is rejected. Now, which differences contributed to the result? Did the respiration caused by the addition of straw differ from that caused by the addition of slurry? Did the kind of straw affect the result? How did the additions of these organic materials affect the respiration rate in relation to the control? These are the preplanned questions that the scientist might reasonably have had in mind when the experiment was designed, and the design should have been such as to answer those questions and test the hypotheses underlying them by the appropriate analysis.

Why preplanned questions? With five different treatments there are 10 different comparisons that can be made between pairs of treatments, and there are more comparisons between combinations of treatments. One might test a comparison between the means of two treatments with a t-test. The standard error for the difference between two treatment means is $\sqrt{2W/n_2}$, so the test is easy to do. Indeed, for the simple balanced case with n_2 replicates per treatment one may compute the least significant difference for comparison between any pair: $\text{LSD} = t\sqrt{2W/n_2}$. With so many possible comparisons it is likely that some will appear 'significant' purely through random variation, and with the human eye and brain well adapted to pick out large differences in tables of means, any inference out of these multiple comparisons is unlikely to be safe. Lark (2017) and Webster (2007) have discussed this matter in greater depth. The meaning of the P-value for a null hypothesis holds when the comparison is planned at the outset; it does not hold for examination of differences after one has inspected a table of means and noted ones that look interesting.

Preplanned questions can be expressed conveniently as a set of orthogonal contrasts. A contrast is a comparison between two treatments, or two groups of treatments. In the example above one contrast might be between soils receiving cattle manure and those receiving pig manure. If we consider the treatments in order control, pig manure, cattle manure, barley straw, wheat straw, then the contrast mentioned can be expressed by a vector of coefficients:

$$\mathbf{c}_1 = [0, -1, 1, 0, 0]\,.$$

This contrast is a comparison between the two manures. There are zero entries that correspond to treatments not in the contrast, and the difference in sign expresses the fact that we are interested in the difference between the two manure treatments.

Another contrast one could consider is between the control and all the treatments with additions to the soil. This would be expressed by the coefficients:

$$\mathbf{c}_2 = [4, -1, -1, -1, -1]\,.$$

Note that the mean for the control has a coefficient of 4, balancing the -1 entry for each of the treatments with an organic amendment, and the coefficients therefore sum to zero, as in the previous example.

We have yet to explain what we mean by an orthogonal contrast. Consider the two examples given. Neither of these contrasts contributes in any way to the other. That is because the second contrast is between the control and all the treatments with an amendment, whereas the first is a contrast between two treatments in the latter group. If we know that the first contrast is large it tells us nothing about the second. Mathematically this is expressed by the fact that the inner product of the two contrast vectors, the sum of the products of their corresponding elements, is zero:

$$\mathbf{c}_1 \bullet \mathbf{c}_2 = 0,$$

as can easily be verified.

We can specify two more contrasts, $\mathbf{c}_3$ and $\mathbf{c}_4$, such that the full set is mutually orthogonal. These are

$$\mathbf{c}_3 = [0, 0, 0, -1, 1]\,,$$

and

$$\mathbf{c}_4 = [0, -1, -1, 1, 1]\,.$$

The contrast $\mathbf{c}_3$ is between wheat straw and barley straw, and the contrast $\mathbf{c}_4$ is between straw and manure. The reader can check that any pair of contrasts drawn from the set $\{\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3, \mathbf{c}_4\}$ is orthogonal.

Note that there are four orthogonal contrasts in this set, which is complete: no additional contrast could be found that is orthogonal to all in this set of four. The number of orthogonal contrasts among a set of treatments is equal to the treatment degrees of freedom. In fact, the orthogonal contrasts can be put into the ANOVA table, one line each, in place of the treatment effects. The treatment sum of squares is partitioned between the contrasts exactly, and each has one degree of freedom. Each contrast can be tested by the ratio of its mean square to the appropriate residual mean square in the design. Note also that orthogonal contrasts can be used in the analysis of a factorial experiment, in which case contrasts can be examined between groups of levels of each factor, and the interaction sum of squares may also be partitioned into corresponding components, each with one degree of freedom.

The use of orthogonal contrasts is much to be commended. It requires experimenters to think in advance about their hypotheses, to express them in terms of contrasts and so to embed them in the experimental design. By prespecifying the orthogonal sets of contrasts experimenters ensure that the P-values they use to test their hypotheses can be interpreted validly.

Often investigators notice, at the end of an experiment, contrasts of interest that they had not expected and for which their design did not cater. Should they apply tests for them? The short answer is 'no'; the only safe way to test the hypothesis implied by such a contrast is to design a new experiment for the purpose.

Several methods have been proposed to test all comparisons post hoc. They include Scheffés critical difference, the Newman–Keuls test, Tukey's 'honest significant difference' and Duncan's multiple range test. The idea underlying them is that by setting the critical limit of P according to the total number of possible comparisons, one can identify which specific contrasts can be regarded as significant. Numerous papers submitted to the journal contain results of these methods to test all comparisons between treatment means, and authors then express the results by littering bar charts or tables of treatment means with letters such that all means with the letter 'a' appended cannot be regarded as significantly different, and so on. This is poor practice. It is of the essence of experimental science to advance hypotheses and to test them; that is the scientist's responsibility. It cannot be delegated to an algorithm. Furthermore, the practice wastes the statistical power of a well-designed experiment, which is only fully exploited by the proper analysis of a set of orthogonal preplanned contrasts. That is why, with the backing of two of the most experienced statistical analysts of the last century, Nelder (1971) and Finney (1988), and the allegorical exposition by Carmer & Walker (1982), this journal eschews routine multiple comparisons from tests.

Nevertheless, these tests can have merit if they are used in what we might call the 'wash-up' phase of the experimental analysis after the primary hypotheses have been tested. They may be used legitimately to 'screen' differences and help investigators to decide whether further research is warranted and to design new experiments accordingly.

In summary, good scientific practice identifies a set of hypotheses that can be expressed as particular preplanned contrasts between the mean responses of treatments or groups of treatments. This is part of the experimental design. The analysis fits the design when the ANOVA table includes the specific orthogonal contrasts as single lines, with one degree of freedom for each mean square, to be tested against the correct residual mean square given constraints on randomization of the treatments between units. If other contrasts catch the experimenter's eye then some of the 'post-hoc' tests listed above might be invoked to screen them.

Some thoughts on sampling

In this paper we have focused on the designs of experiments and the analyses of variance for inference from data obtained according to those designs. Similar considerations apply to sampling to estimate, for example, the mean values of soil properties within regions of interest. We have described suitable designs elsewhere (Webster & Lark, 2013), and we cannot go into detail here. Readers can find the general principles in the classic text by Cochran (1977) and their application to spatial sampling in De Gruijter *et al.* (2006).

In sampling, as with experiments, the principle that the analysis should fit the design still holds good. In the context of sampling our objective is estimation, and an estimate should be accompanied by a confidence interval to indicate its precision. There are standard methods to compute such confidence intervals, but the method that is used must accord with the sampling design if it is to be safe. For example, most soil scientists would recognize the procedure of computing the sample variance, s^2, from a set of N observations and then calculating the standard error of the sample mean as

$$\frac{s}{\sqrt{N}} \ . \tag{9}$$

One can compute the confidence interval for the sample mean by multiplying the standard error by the value of Student's t for which the distribution function with $n-1$ degrees of freedom takes an appropriate value (e.g. 0.975 for the 95% confidence interval). This simple analysis is appropriate, however, only when the N samples have been collected independently and completely at random (also known as simple random sampling). Without the independence, which independent random sampling ensures, the computation of the standard error in Equation (9) is wrong.

Too often the journal receives papers in which the analysis of sample data does not fit the design. Most commonly that is because the authors use Equation (9) to compute the standard error of a sample mean based on N samples that were not collected independently and at random, either because the sampling was not randomized (sample sites may have been selected purposively to cover a range of soil variation) or because the samples were collected according to a systematic design (a grid or transect). In the latter, once the positions of one or two sampling sites have been chosen the positions of all the others in the designs are determined by the interval of the grid or transect. One may compute a correct standard error for an estimated mean where sampling has been done systematically on several transects provided the starting points of the transects are chosen at random (De Gruijter *et al.*, 2006) and the analysis fits the design appropriately. Alternatively, model-based estimation may be used (Lark & Cullis, 2004).

Other sampling designs may be appropriate. Stratified random sampling is directly analogous to the RCB experimental design discussed above. The domain of interest is divided into strata, which one hopes are less variable internally than the domain as a whole. The estimates are likely to be more precise than those from simple random sampling because the estimation variances are based on the variances within the strata rather than on those of the whole domain. Each stratum is sampled independently and at random, the stratum sample means are combined to obtain an estimate of the domain mean, and the stratum variances are similarly combined to obtain a variance of the estimated mean. If stratification has been used in the sampling design then it must be accounted for in the analysis.

Departures from assumptions

We have stressed throughout that the correct analysis of variance fits the design; no other will do. The conclusions that you may draw from such analyses, however, are based on the assumption that the effects of the various factors (treatments and blocks and their combinations) are additive, that the residuals are normally and independently distributed, and that the variances are homogeneous. Small departures from these ideal conditions are unlikely to affect your conclusions: the analysis of variance is robust in this respect. Large ones, on the other hand, might. Testing for serious departures and the transformations required to make data conform to the assumptions are substantial subjects in their own right, and we cannot deal with them here. Instead we refer you to Chapter 15, pages 273–296, in Snedecor & Cochran (1989), and Chapter 8, pages 159–181, in Mead *et al*. (2003).

Epilogue

This paper is not a comprehensive account of the design and analysis of experiments; it was never our intention that it should be. Rather, we have wanted to stress the importance of sound experimental designs, of doing experiments according to those designs and then subsequently analysing the data that accrue. Readers can find details of the designs we mention in the texts we have cited; those texts should cover their requirements.

Sound inferences about the effects of treatments on the soil demand that treatments are replicated and assigned to experimental units at random. The natural variability of the soil is substantial, and many replicates might be needed to reveal the effects of the treatments against this backdrop of natural variation. One can often reduce the amount of replication, and increase the efficiency of an investigation, by blocking. Whether a completely randomized design is used, or a randomized complete block design, the design must be accounted for in the analysis, and it should be made explicit by the full ANOVA table. If your paper does not contain such a table then readers cannot be sure that you have analysed your data in a way that fits the design and is therefore valid.

More complex experimental designs might be needed for practical reasons. We have given the example of split plots, but others include designs with incomplete blocks and designs in which certain interactions are deliberately confounded and so cannot be estimated. In all cases the experimental design constrains the analysis, and the degrees of freedom in the ANOVA table, and the residual mean square against which an effect is tested, must accord with the design as described. The same holds for repeated measures on the same experimental units, and for experiments when replicated samples from within the basic experimental units are analysed separately.

Finally, we have stressed that scientists have the responsibility to propose hypotheses and to design experiments accordingly. By preplanning particular comparisons scientists embed their hypotheses in those designs. Their analyses partition the treatment sums of squares into components corresponding to the orthogonal contrasts.

Soil scientists nowadays use some of the most advanced techniques from nuclear magnetic resonance to shallow geophysics, and we like to think that they take advice from specialists beforehand. They should do the same when they apply statistical methods. Modern software provides a wide range of readily available tools for statistical analysis. But when misused by investigators who lack proper understanding they lead to flawed inferences, and those can have damaging consequences if they lead in turn to bad decisions by farmers, environmental managers, statutory authorities and agencies responsible for public health.

We encourage soil scientists to think hard about how they design their experiments and then analyse the data. We encourage educators in soil science to ensure that statistics, taught by specialists, has an essential place in curricula at both the undergraduate and postgraduate level. Finally, we urge soil scientists to consult statisticians when they plan their experiments, and not go along to them at the end and ask them how to analyse their data. Neither you nor we want Fisher to look down and pronounce yet another post-mortem on your experiment.

Supporting Information

The following supporting information is available in the online version of this article:

File S1. As mentioned above, we have provided examples of completely randomized (CR), randomized complete block (RCB) and split-plot designs with simulated data, together with programs in GenStat and R for the correct analyses of variance and the output from those analyses. This supporting information in files exp1.csv, exp1.gen, exp1.R, exp1.outGS.txt, exp1.outR.txt, exp2.csv, exp2.gen, exp2.R, exp2.outGS.txt, exp2.outR.txt, exp3.csv, exp3.gen, exp3.R, exp3.outGS.txt, exp3.outR.txt and readme.txt, is available in the on-line version of this article and also from us as the file Supplementary material.zip.

References

Carmer, S.G. & Walker, W.M. 1982. Baby Bear's dilemma: a statistical tale. *Agronomy Journal*, **74**, 122–124.

Coban, H., Miltner, A., Centler, F. & Kästner, M. 2016. Effects of compost, biochar and manure on carbon mineralization of biogas residues applied to soil. *European Journal of Soil Science*, **67**, 217–225.

Cochran, W.G. 1977. *Sampling Techniques*, 3rd edn. John Wiley & Sons, New York.

Cochran, W.G. & Cox, G.M. 1957. *Experimental Designs*, 2nd edn. John Wiley & Sons, New York.

De Gruijter, J.J., Brus, D.J., Bierkens, M.F.P. & Knotters, M. 2006. *Sampling for Natural Resources Monitoring*. Springer-Verlag, Berlin.

Finney, D.J. 1988. Was this in your statistics textbook? III. Design and analysis. *Experimental Agriculture*, **24**, 421–432.

Fisher, R.A. 1926. The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain*, **33**, 503–513.

González-Méndez, B., Webster, R., Fiedler, S., Loza-Reyes, E., Hernández, J.M., Ruíz-Suárez, L.G. *et al.* 2015. Short-term emissions of CO_2 and N_2O in response to periodic flood irrigation with waste water in the Mezquital Valley of Mexico. *Atmospheric Environment*, **101**, 116–124.

González-Méndez, B., Webster, R., Fiedler, S. & Siebe, C. 2017. Changes in soil redox potential in response to flood irrigation with waste water in central Mexico. *European Journal of Soil Science*, **68**, 886–896.

Jeffers, J.N.R. 1978. *Design of Experiments.Statistical Checklist 1*. NERC Institute of Terrestrial Ecology, Grange-Over-Sands [WWW document]. URL http://nora.nerc.ac.uk/5271/ [accessed on 21 July 2017].

Kenny, A.J. 2005. *Wittgenstein*. Wiley–Blackwell, Oxford.

Kenward, M.G. 1987. A method for comparing profiles of repeated measurements. *Applied Statistics*, **36**, 296–308.

Lark, R.M. 2017. Controlling the marginal false discovery rate in inferences from a soil data set with α-investment. *European Journal of Soil Science*, **68**, 221–234.

Lark, R.M. & Cullis, B.R. 2004. Model-based analysis using REML for inference from systematically sampled data on soil. *European Journal of Soil Science.*, **55**, 799–813.

Mead, R., Curnow, R.N. & Hasted, A.M. 2003. *Statistical Methods in Agriculture and Experimental Biology*. Chapman and Hall/CRC, Boca Raton, FL.

Nelder, J.A. 1971. Discussion on papers by Wynn, Bloomfield, O'Neill and Wetherill (1971). *Journal of the Royal Statistical Society B*, **33**, 244–246.

Rothamsted Research 2006. *Guide to the Classical and Other Long-Term Experiments, Datasets and Sample Archive*. Lawes Agricultural Trust, Harpenden.

Snedecor, G.W. & Cochran, W.G. 1989. *Statistical Methods*, 8th edn. Iowa State University Press, Ames, IA.

Webster, R. 2007. Analysis of variance, inference, multiple comparisons and sampling effects in soil research. *European Journal of Soil Science*, **58**, 74–82.

Webster, R. & Lark, R.M. 2013. *Field Sampling for Environmental Science and Management*. Routledge, London.

Webster, R., Oliver, M.A. & Lark, R.M. 2016. Editorial: statistics in the journal. *European Journal of Soil Science*, **67**, 133–134.

Webster, R. & Payne, R.W. 2002. Analysing repeated measurements in soil monitoring and experimentation. *European Journal of Soil Science*, **53**, 1–13.

Wilkinson, G.N. & Rogers, C.E. 1973. Symbolic description of factorial models for analysis of variance. *Applied Statistics*, **22**, 392–399.

Yates, F. 1937. *The Design and Analysis Of Factorial Experiments*. Technical Communication 35. Commonwealth Bureau of Soil Science, Harpenden.