SOME ASPECTS OF THE ANALYSIS AND PRESENTATION OF DATA FROM COMPARATIVE TRIALS

N.R. COX

Biometrics, Ruakura Agricultural Research Centre, MAF, Private Bag, Hamilton

SUMMARY

Three examples are used to illustrate aspects about the analysis and presentation of data from comparative trials typical of those presented at this conference. The meaning, assumptions and interpretation of the analysis of variance is discussed. The estimation aspect is stressed, that of significance testing down-played. Means, an appropriate standard error and preferably no Duncan's letters or similar, should be presented in tables while the text should describe the important trends or differences.

INTRODUCTION

To illustrate the various aspects of analysing and interpreting the results of experiments typical of those reported in this publication, three examples will be used.

Aphid trial

This trial is being reported at this conference by Wrenn and Skinner (1982). The experiment involved spraying lucerne with two different aphicides each at three water volumes using a controlled droplet applicator plus one volume using a conventional boom. These eight combinations were applied in a randomised block design in five replicates.

Takapau grass grub trial

The results of trials at Takapau and Paewhenua were reported by Pottinger *et al* (1977) in the proceedings of the 30th conference of this society. Cypermethrin was used with a sex attractant (to determine application time) for controlling grass grub. For the Takapau trial, one, two or three sprays were used at each of four rates in a randomised block design in four replicates.

Barley variety trial

Duncan (1955) in his paper that describes his well-known multiple range test uses a barley variety trial for illustration. Seven varieties were tested replicated six times, again in a randomised block design.

ANALYSIS OF VARIANCE

The first step in the analysis of the data obtained from each of these trials would be to carry out an analysis of variance to define the influences of the various treatment effects.

All we are doing in an analysis of variance is setting up a linear model to describe the data and using the method of least squares to obtain estimates of the unknown effects. The total variation in the data is also partitioned into components associated with these effects. This enables us to compare the sizes of the various contributions to see what evidence there is that a treatment is having a real effect and to determine the size of the residual unexplained variation.

While the method of least squares requires no assumptions about the distributions of the data, the use of the associated F ratio (variance ratio) tests, t-tests and confidence intervals does imply certain assumptions. The assumptions required are that the observations,

- (i) are normally distributed
- (ii) all have the same variance, and
- (iii) are independent of each other.

One should not do an analysis of variance without checking these assumptions.

Proc. 35th N.Z. Weed and Pest Control Conf.

Fig. 1: (Top left) Residuals from analysis of variance of untransformed aphid counts versus fitted values and normal scores.

Fig. 2: (Bottom left) Residuals from analysis of variance of log transformed aphid counts versus fitted values and normal scores.

Fig. 3: (Right) Ordered barley variety yield means versus normal scores.

To show how to check these assumptions and to overcome non-compliance with them, the aphid counts recorded 7 days after spraying in the aphid trial will be used. The first graph in figure 1 shows plots of residuals against the corresponding fitted values. The residuals are the unexplained differences between the observed data values and the values (fitted values) predicted from the linear model. A wedge shape is obvious in this, showing that the variability is increasing as the fitted values increase and hence that assumption (ii) is violated.

In the second graph the ordered residuals are plotted against the normal scores. The meaning of normal scores will not be explained here, though in the next section there is some brief discussion. It is sufficient to know that the points in the second plot would lie roughly along a straight line if the residuals were from a normal distribution. The S-shaped appearance here shows that the normality assumption is violated. The analysis of variance on this data is therefore invalid.

However it is often true that some transformation of the raw data will meet the analysis of variance assumptions. For this and other reasons the log transformation was tried. The residual plots from this analysis are shown in figure 2. It can be seen at once that the wedge shape has disappeared in the first graph and the line in the second graph is very nearly straight. The analysis of variance of the log-transformed data is therefore valid.

The third assumption, that of independence, is perhaps the most critical in the sense that lack of independence can drastically mislead. Had there been large differences between the blocks in this experiment and had the block effect been omitted from the analysis of variance model, the dependencies between the residuals would have been apparent. So the omission of an important effect from the model is one way in which the independence assumption can be violated.

Poorly designed experiments can generate dependencies. For instance in this experiment "contamination" of neighbouring plots through spray drift or aphid migration could have generated spacial dependencies. Sometimes such carry-over effects cannot be avoided and need to be taken account of in the analysis of the data in order that problems associated with lack of independence be avoided.

It was stated above that the log transformation was used for reasons additional to the fact that it generated normality and constant variance. It is reasonable to suppose that insecticide treatments act in a proportional way; e.g. a chemical may kill 90% of a population irrespective of the density of that population rather than killing a fixed number of insects. By taking logs before fitting the (additive) analysis of variance model, we are actually fitting a model that is multiplicative (or proportional) on the original data. In other words, we are fitting a model that is biologically much more reasonable. We should never lose sight of the fact that we are, in an analysis of variance, attempting to describe the data by means of a simple model and therefore we should think very carefully about whether or not the model we are fitting is sensible.

INTERPRETATION

The analysis of variance of the day 7 (logged) aphid counts showed a highly significant aphicide by rate interaction. The means and the standard error of differences are shown in Table 1.

TABLE 1: Means of logged aphid counts

Insecticide	Conv. 300	Volume (litres/ha)			Mean
		CDA15	CDA30	CDA45	
fenvalerate	2.2	5.6	0.9	0.4	2.3
thiometon	0.3	4.6	4.2	1.9	2.7
average	1.3	5.1	2.5	1.1	2.5

se (differences) = 0.53 (for values in body of table)

The establishment of the significant interaction is only the first step. Next the reason for the interaction should be investigated. In this case inspection of Table 1 shows (amongst other things) that for fenvalerate the two higher rates using CDA were the best while for thiometon the conventional method gave the best results, the higher CDA volumes being much less effective than with fenvalerate.

The standard error shown in Table 1 is most important. It is calculated from the residual variation in the analysis of variance. This residual variation indicates the size of the chance or unexplained variation in the data and the standard error given in the table gives a measure of the accuracy with which differences have been estimated.

The estimated difference between the 15 and 30 litres/ha volumes for thiometon is 0.4; the standard error of 0.53 shows that this difference could easily be fortuitous. Differences greater than about 2 standard errors are unlikely to occur by chance and so the difference between the chemicals at the 45 litres/ha volume, i.e. 1.5, is strong evidence of a real difference.

Given the significant chemical by volume interaction, the question of an overall difference between the two chemicals does not arise. For instance, one cannot say simply that fenvalerate is better than thiometon; it is for the higher volumes using CDA but not for the lower volume or for the conventional boom.

When there is no real evidence of an interaction we can go on to examine possible differences between main effect levels. In the barley variety trial with only one main effect, the analysis of variance of the grain yields showed a highly significant variety effect indicating strong evidence of real differences between the varieties. The means are shown in Table 2.

As with the examination of the causes for the interaction in the aphid trial, we should look closely at the means to see just what is giving rise to the significant effect.

TABLE 2: Barley grain yields

Variety	A	F	G	D	C	B	E
Yield	49.6	58.1	61.0	61.5	67.6	71.2	71.3
Duncan's MRT (5%)	c	bc	abc	ab	ab	a	a

Se (differences) = 5.15

Perhaps one mean is discrepant from the rest or perhaps the means fall into two or three distinct groups. A good aid to discovering such patterns is to plot the observed means against the normal scores (also called the expected normal order statistics). These scores are easily generated on a computer and tables are also available, e.g. Fisher and Yates (1963). They indicate what spread of observed means you could typically expect when there is no difference in the true means.

The plot for Duncan's example is shown in figure 3. Also shown in the figure is a line drawn through the overall mean and with slope equal to the standard error of the means. If there were no real differences between the varieties, the points would lie pretty much along this line. It is easy to see from this graph the main discrepancies from the pattern that would be expected with no real differences. There is an unexpectedly large gap between the first three and the next three and between these and the last one. These groupings are far from clear from the Duncan's letters in the table which certainly do not hint at such groupings. The evidence is that E, B and C are better than the rest while A is worse than all others.

In my opinion that is all that is required, with the means and standard error presented in a table and a general interpretation in the text. A reader interested in a particular comparison can see for himself how strong the evidence is for that difference. For example, the B-F difference is more than twice the standard error of the difference and so there is very good evidence that B is better than F. On the other hand the C-D difference is little over one standard error and so the evidence in favour of C being better than D is not so strong even though the combined evidence is that C is amongst the best and D is not; the situation is not clear cut.

ESTIMATION

In my experience most researchers do experiments not so much to see whether or not something has an effect but rather to see *how much* effect it has. For instance, in field applications of insecticides, the experimenter will be using chemicals that he already knows will kill insects (at least under laboratory conditions) and so in his field experiments he wishes to determine just *how* good it is. And when more than one rate of insecticide is applied it is really not a question of whether a high rate kills more than a lower rate but of *how many* more the high rate kills.

The situation is much the same in fertiliser and herbicide trials, nutritional studies and so on. I find in reading experiment project proposals at Ruakura that in most cases there is quite a lot of information already available, perhaps only of a qualitative nature, and that the objectives of the projects are to quantify the effects.

In the grass grub trial, for instance, some laboratory and field work had already been done and the reported trial was part of a project that had as its objective the optimisation of the treatment. Therefore the experiments were aimed at determining the level of control achieved by various rates, the additional control achieved by second and third applications and so on. I am sure Dr Pottinger had little doubt that if he sprayed more than once he was bound to kill more grass grubs. So a test of the hypothesis that two sprays is equal to one spray is not really relevant. What he wants to know is how many more the second spray kills so that he can see if the additional cost is going to be worth it.

In the table of results reproduced below (Table 3), however, the equality of all pairs of rate and spray combinations is compared by means of Duncan's test. The apparently complex conclusions about which treatment is better than which other treatment are quite irrelevant. In reality the conclusions appear quite straightforward, though I have

TABLE 3: Effects of cypermethrin for control of grass grub in pasture

Rate (kg/ha)	No. of sprays	3rd-instar larvae/m ²
0.60	1	151.5 abc
0.30	1	57.3 e
0.15	1	140.0 abcd
0.075	1	104.4 bcde
0.60	2	108.2 b
0.30	2	120.9 abcde
0.15	2	187.1 a
0.075	2	95.5 cde
0.60	3	50.9 e
0.30	3	70.0 de
0.15	3	72.5 de
0.075	3	85.3 cde
untreated	-	170.0 ab

not analysed the data properly myself, and I am pleased to say that the text does clearly state these simple conclusions. Essentially, for the Takapau data, three sprays gave 60% control while one or two sprays gave only 33% control.

CONCLUSIONS

For the type of experiment of which the above examples are typical, the first step should be an analysis of variance, checking that the assumptions are justified and transforming if necessary. This will indicate which effects or combinations of effects have had an important bearing on the results. From the residual variation, standard errors for differences between means can be calculated.

The pattern of means for significant effects should be examined carefully to discover the cause of the significant effect and estimates of relevant differences or trends obtained.

In reporting their results, authors should display relevant estimates or means and quote corresponding standard errors. I see no place for multiple range tests such as Duncan's test; in many cases the test is inappropriate or irrelevant as I showed for the grass grub trial and the letters are a poor aid for understanding differences, as Duncan's own example shows.

For the grass grub trial, for instance, the means should be displayed along with a standard error and minus any Duncan's letters. In the text, the important patterns should be described, in this case that three sprays gave 60% control while one or two sprays gave only 33% control.

ACKNOWLEDGEMENT

Thanks go to Mr N.R. Wrenn for supplying the data for the aphid trial.

REFERENCES

- Duncan, D.B., 1955. Multiple range and multiple F tests. *Biometrics* 11: 1-42.
 Fisher, R.A. and Yates, F., 1963. Statistical tables for biological, agricultural and medical research, 6th Edition, Longman, London: p.94.
 Pottinger, R.P., East, R., Welsh, R.D., Stratton, A.E. and Maxwell, B.C., 1977. Evaluation of cypermethrin for control of adult grass grub. *Proc. 30th N.Z. Weed and Pest Control Conf.*: 217-21.
 Wrenn, N.R. and Skinner, R.A., 1982. The effect of droplet size and spray-mix volume on control of blue green lucerne aphid. *Proc. 35th N.Z. Weed and Pest Control Conf.*: In press.