Expl Agric. (2000), volume 36, pp. 343–351
Printed in Great Britain
Copyright © 2000 Cambridge University Press

ABERRANT DATA IN FIELD EXPERIMENTS

By G. V. DYKE† *and* S. C. PEARCE‡

†*Rothamsted Experimental Station, Harpenden, Herts AL5 2JQ, UK, and*
‡*Institute of Mathematics and Statistics, University of Kent at Canterbury, Canterbury, Kent CT2 7NF, UK*

(*Accepted 11 January 2000*)

SUMMARY

A field experiment will sometimes give data so unexpected that their correctness is doubted. It would be wrong to discard them because they go against the experimenter's preconceptions but sometimes enquiry will reveal why they are aberrant. Sometimes no reason can be found but they still raise doubts. The problem is approached using a published body of data.

The calculation of residuals and total residuals is commended as a means of enquiry. Exclusion of data is to be avoided if possible. It can have serious consequences and methods are given for assessing them. It is important to seek alternative approaches and one possibility is the use of covariance adjustments on fertility trends shown by study of the residuals. It is shown also that extreme values sometimes arise from unsuitable data and can be eliminated by a suitable transformation, but it must be meaningful and not just a statistical expedient to avoid difficulties.

INTRODUCTION

The problem

However careful an investigator may be in the conduct of a field experiment, some of the recorded data may raise doubts. It could be that the fertility pattern of the field was misjudged and in consequence an inappropriate design was chosen or perhaps some forgotten event in the past had left its mark on the site. The trouble may not derive from the land at all but from an unforeseen mishap or a fault in field technique.

Given suspect data, the experimenter will naturally want to salvage as much information as possible and to do so will have to conjecture what might have gone wrong. This paper will be illustrated by the problem of Basford and Tukey (1995) in which one plot yielded much more than was expected and a possible explanation was discovered. Their paper exhibits the effective use of two little-known techniques, namely 'median polishing' and 'Winsorization'. Here it will be argued that the procedure should depend upon the investigator's assessment of the reason for some data being aberrant and various simpler possibilities will be suggested.

In the present study the reader would be helped by a previous knowledge of

https://doi.org/10.1017/S0014479700003057 Published online by Cambridge University Press

Table 1. Plan, data, total residuals and residuals for the experiment used as an example.

Block			*Treatments*				
1	*	4	3	2	*	1	
2	*	3	1	4	2	*	
3	3	2	4	1	*	*	
4	1	2	3	4	*	*	
			Data				
1	*	419	193	71	*	43	
2	*	230	108	41	78	*	
3	278	206	103	42	*	*	
4	160	99	45	44	*	*	
			Total Residuals				Mean
1	*	+267	+7	−43	*	−45	+46
2	*	+44	+20	−111	−36	*	−21
3	+92	+92	−49	+2	*	*	+22
4	+56	−15	−143	−93	*	–	−48
Mean	+82	+97	−41	−77	−36	−45	
			Residuals				
1	*	+221	−40	−89	*	−92	
2	*	+65	+40	−71	−69	*	
3	+70	+70	−71	−69	*	*	
4	+120	+34	−95	−59	*	*	

s.e. of each residual = 89.4

contrasts (Pearce, 1992) and strata (Pearce, 1995). There are many other texts, but the two cited are in this journal and therefore readily accessible.

The data are set out in Table 1. In full the experiment, which continued for several years, had six plots in each of four blocks, but in the season considered only four of the treatments carried sweetcorn (*Zea mays*). (In this season two treatments had not yet differed, but Basford and Tukey ignore this in order to give a simple example.) In consequence the data used are from only four treatments, which form a randomized block design. The data represent yields of sweetcorn cobs, expressed as kg per plot multiplied by 10. That for Treatment 4 in block 1 is markedly out of line with the others.

Residuals and total residuals

In a search to discover what may have gone wrong it is sound practice to calculate residuals and to enter their values on a copy of the experimental plan as in Table 1. For a design in randomized blocks the calculations are very simple. First each datum is reduced by the appropriate treatment mean to give the 'total residual', which estimates the inherent fertility of the plot and so draws attention to trends. In the present instance there is clearly such a trend from the top left to the bottom right. Then, taking the total residuals, each is reduced by the

https://doi.org/10.1017/S0014479700003057 Published online by Cambridge University Press

appropriate block mean to give the 'residual' for the plot. The residual sum of squares in the analysis of variance equals the sum of the squared residuals. Consequently knowledge of their individual values may show that the greater part of the error comes from a limited area or from only a few treatments. In this instance it is clear that the suspect datum is making much the largest contribution.

The table of residuals so formed should be completed by an estimate of the standard error applicable to each residual. It is the square root of RSS/N, where RSS is the residual sum of squares and N is the number of plots in the experiment. It needs to be used with caution because residuals of plots are subject to correlations, some positive and some negative. (Since residuals may be needed in subsequent calculations it is advisable to carry more decimal places than are needed for purposes of inspection.)

EXCLUSION OF DATA

Effects of exclusion

Some experimenters feel that the best way of dealing with dubious data is simply to exclude them, but care is needed. Admittedly there is usually little technical difficulty about analysing an incomplete set of data, but the omission of data can have serious effects and may result in such loss of vital information as to make analysis of the remaining data of little value. There is a more important reason however. Here the plot exists and a value (419) was found for it; it should not be discarded just because it does not fit someone's preconceptions. An experiment must not be subjected to censorship solely because it does not support an existing line of thought.

In the example of Basford and Tukey there are four blocks each of which contains four treatments. In one of them Treatment 4 has been discarded. Consequently that block cannot be used to estimate the values of some of the contrasts which may be of interest. For example, the contrast $C_1 = (+1\ +1\ +1\ -3)$, will have to be based on three replicates, not four. On the other hand contrasts like $C_2 = (+1\ -1\ 0\ 0)$ and $C_3 = (+1\ +1\ -2\ 0)$ are unaffected and are still estimated as for complete data. If the object of the experiment is to compare three fertilizer treatments, 1, 2 and 3, with one another and the fourth treatment with no added fertilizer has been included to provide a base line, the loss of information about C_1 might not matter much, but with a different treatment structure it could be more serious. The importance of loss of data can be assessed only by reference to the structure of the treatments.

A general rule

A general expression for the effect of lost data on contrasts has been given by Burns *et al.* (1978). If a datum is missing, let C_0 represent the contrast between the damaged treatment and all the others. Also let C_x be the contrast under study. If now ρ is the correlation coefficient between the elements of the two contrasts, the loss of data represented by C_0 reduces the effective replication of C_x by at least ρ^2.

https://doi.org/10.1017/S0014479700003057 Published online by Cambridge University Press

In our case C_0 happens to equal C_1 defined above.The correlation coefficient between the elements of C_0 and C_2 equals 0 so the damage represented by C_0 does not affect the estimation of C_2. Neither does it affect C_3. On the other hand, C_0 and C_1 are the same so $\rho^2 = 1$ and the effective replication of C_1 is reduced by that amount. That does no more than confirm what has already been concluded but the method can be applied more widely. If the treatments had formed a 2×2 factorial set with two main effects, $(+1\ +1\ -1\ -1)$ and $(+1\ -1\ +1\ -1)$, and their interaction $(+1\ -1\ -1\ +1)$, then for each of the contrasts of interest ρ^2 would have been 1/3 and the best effective replication that could be hoped for is 3.67. For all three the value is in fact 3.60, so the loss of one-sixteenth of the data has led to the loss of one-tenth of the desired information. It is true that some can be recovered from the inter-block analysis, but not very much.

Doubtful plot values

Although it is wrong to exclude a datum just because it looks wrong there are occasions when examination shows the residual to be so large as to raise serious doubts as to some unidentified mishap. Also sometimes, as in the example of Table 1, there is in fact a reason for questioning the aberrant value. It had been discovered that the plot might have been affected by the dumping of rock phosphate 25 years earlier. If it had been, some of the adjacent plots could have been affected to a less degree and there could well be residues detectable by soil analysis. No doubt suitable enquiry was made and those with local knowledge believed that they had found a possible reason for the aberrant datum.

The matter can readily be tested. Analysing all 16 data gives a residual sum of squares equal to 127 994 with nine degrees of freedom. If the aberrant value is excluded the sum of squares is 41 362 with eight degrees of freedom, giving a mean square of 5170.2, so the contribution from that one plot is 86 632 (= 127 994 −41 362) with one degree of freedom. That leads to a variance ratio of 16.76 (= 86 632/5170.2) with one and eight degrees of freedom, a value that would arise by chance about once in 200 times and this implies that the datum in question had indeed been subject to some special disturbance.

There is a simple way of carrying out the required calculation (Rubin, 1972). First a pseudo-variate is required with value 1 for the questionable plot and 0 for all the others. Let $\mathcal{Z}$ denote the residual sum of squares in an analysis of variance of the pseudo-variate.(For an orthogonal design $\mathcal{Z}$ simply equals f/N, where the full design has N plots and f residual degrees of freedom.) Then the decrease in residual sum of squares brought about by discarding the doubtful plot will be $\mathcal{Z}e^2$, where e equals the actual datum less its estimated missing-plot value. In our example, this estimated value is found to be 26.556 by the usual method, so that e equals 419 − 26.556 = 392.444. Also, $\mathcal{Z} = 9/16$, whence $\mathcal{Z}e^2 = 86\,632$, the value given above.

Alternatively, a valid analysis of variance for the incomplete data can be found by putting the doubtful value equal to zero and carrying out an analysis of covariance in which adjustment is made on the pseudo-variate (Bartlett, 1937).

https://doi.org/10.1017/S0014479700003057 Published online by Cambridge University Press

The missing plot value is no longer required but it is in fact $-b$, where b is the regression coefficient. These methods have been discussed at greater length elsewhere (Pearce, 1983).

So far attention has been given only to the stratum of plots within blocks (Pearce, 1995) but the loss of a datum has transferred a small amount of information concerning treatments to the stratum of blocks within the whole. The one block that no longer contains Treatment 4 can be compared with the three that still have it. Use of residual maximum likelihood (REML), which uses information from all strata of a data set with appropriate weighting (Patterson and Thompson, 1971) gives an estimate of within-block variance of 5106, only a little less than the figure already obtained (5170). It leads to slightly smaller standard errors of differences (Table 3). Rounded to the nearest unit, only two of the residuals differ by ± 1 from those found from the analysis of variance of 15 plots.

ALTERNATIVES TO OMISSION

Covariance adjustments

Study of the residuals and total residuals may force the unwelcome conclusion that there is some feature of the land that was not taken into account when choosing the design. There could, for example, be an area of poor growth, which on investigation turned out to correspond to some measurable characteristic such as poor soil pH or depth. It is not possible to change the design because the use of blocks (or rows and columns) has led to a restriction of the randomization that must not be ignored, but a remedy can often be found by using the analysis of covariance to adjust for the disturbing feature.

A similar situation arises here. Examination of the total residuals in Table 1 suggested that there was a trend across the site from top left to bottom right so marked that it cannot be dismissed as coincidental. If blocks had not been inserted it would have been reasonable to adjust for this trend using the pseudo-variate (x_1) on the left side of Table 2. Since blocks do exist and allow for any trend from top to bottom, it is more rigorous to use the pseudo-variate (x_2) on the right for the remaining component of the trend, that is, that from left to right.

Table 2. Values of pseudo-variates that would allow adjustment for the fertility trend using the analysis of covariance.

x_1						x_2					
*	5	4	3	—	1	*	5	4	3	—	1
*	4	3	2	1	*	—	5	4	3	2	—
4	3	2	1	*	*	6	5	4	3	—	—
3	2	1	0	*	*	6	5	4	3	—	—

x_1 represents slope parallel to diagonal of plot; x_2 represents trend from left to right

https://doi.org/10.1017/S0014479700003057 Published online by Cambridge University Press

Table 3. Summary of alternative analyses of the experiment.

Analysis	Number of plots	Treatment means 1	2	3	4	s.e.d.	Variance ratio	d.f.	d.f. Wald (3 d.f.)
ANOVA	16	88	114	186	152	84.3	0.52	9	—
ANOVA	15	88	114	186	54	50.8, 54.9	2.18	8	—
REML	15	88	114	186	64	50.5, 54.4	—	—	6.1
Covar1	16	117	126	134	164	42.6	0.48	11	—
Covar2	16	118	126	131	165	47.0	0.40	8	—

Notes: ANOVA: analysis of variance. REML: residual maximum likelihood. Covar1: analysis of covariance; covariate x_1, blocks ignored. Covar2: analysis of covariance; covariate x_1, blocks fitted. Wald: Wald test, asymptotically equivalent to χ^2

The calculations can now be repeated after allowing for the suspected trend by adjustment on the covariate. In the new analysis the residual sum of squares is 37 715 with eight degrees of freedom, giving a mean square of 4714 (= 37 715/8), a marked improvement on the former value of 14 760. Also the residual of the dubious plot has been reduced to 87. It appears then that there is a marked fertility trend across the field. It might account for the aberrant value.

To test the matter, the plot in question will be regarded as missing and assigned the value of 27 by the usual formula. The covariate retains its value of five. In the resulting analysis of covariance the residual sum of squares is reduced to 22 547 with seven degrees of freedom. It follows that the aberrant plot value must be contributing 15 168 (= 37 715 − 22 547) with one degree of freedom. Also the mean square is now only 3221 (= 22 547/7), so that the variance ratio for the deviation of the value from its expectation has become 4.71 (= 15 168/3221), a value that would arise by chance about once in 12 times and does not suggest that the figure of 419 is very remarkable.

But, provided that administrative changes (such as planting by different workers or harvesting on different dates) were not confounded with block differences, some analysts may feel justified in using x_1 as covariate and omitting blocks from the analysis, leaving more residual degrees of freedom. Table 3 shows both analyses of covariance.

Transformation of data

Aberrant data sometimes result from the form of distribution of the data, which in this instance seems clearly to be far from Normal. Examination suggests that the data would serve better if they were transformed to their logarithms, but care is needed. The matter cannot be decided on the basis of only 16 data; confirmation is needed, either from past experience or from an argument as to why the transformed data would be better.

A conflict arises because the figures in a report should relate to the needs of the reader, who may find difficulty in interpreting logarithms of yield. On the other hand, data should conform to the assumptions behind the method of analysis and

https://doi.org/10.1017/S0014479700003057 Published online by Cambridge University Press

in this instance they clearly do not. For example, all should be subject to the same variance regardless of treatment, but here the treatments give ranges of 118, 135, 233 and 377 respectively, values that rise even more steeply than the treatment means. Where the ratio of range to mean is approximately constant, the analysis is often improved by analysing not the data but their logarithms. Here something even stronger may be indicated, but caution is needed.

If the data were a measure of size, for example stem length, logarithms could be presented as an indication of growth rate, which is of course the characteristic actually affected by treatments and environmental differences, but that argument cannot be applied here. One possibility is to use transformed data in the statistical analysis and to back-transform means etc by antilogarithms so that the figures presented are in the original units. In practice, this usually has the effect of reducing all means in about the same proportion, which does not distort the relative treatment contrasts but can provoke the objection that the actual means are important in themselves.

Two points of view are possible; first, that the larger data are subject to larger random variation and should therefore carry less weight than smaller data in the determination of means or, second, that the only unbiased estimate of a treatment-contrast over the whole site is the unweighted mean of replicates. It would be idle to pretend that there is one accepted decision. It is clear only that aberrant values sometimes arise because the data do not follow the Normal distribution assumed in statistical theory.

When the data are analysed after transformation to natural logarithms the residual sum of squares is 6.8833 with nine degrees of freedom, the residual of the doubtful datum being +2.30. Its contribution is therefore 2.9845 $(= 2.30^2 \times 9/16)$, leaving 3.8988 with eight degrees of freedom, that is, a residual mean square of 0.4874. Hence the variance ratio is 6.12 $(= 2.9845/0.4874)$ with one and eight degrees of freedom, a figure that may be expected to arise by chance about once in 26 times; this is unremarkable considering that the residual was not chosen at random but as the largest of 16. The point is not that the analysis of transformed data is necessarily correct but that aberrant data can result from an ill-chosen variate. In this instance the assumption that all data are equally subject to error is plausible with the transformed data though perhaps unrealistic with the untransformed.

DISCUSSION

An experiment exists to provide the data required for someone to make a decision, either a way of thinking has been supported or it has been questioned. Once obtained the data are the arbiter. If experimenters do not like what they have recorded they may disregard it only for a good reason. Perhaps someone has discovered a feature in the site that should have been allowed for in the design but was not known about at the time, or perhaps a machine has developed a fault. Perhaps someone can be shown to have made a mistake and some or all of the

https://doi.org/10.1017/S0014479700003057 Published online by Cambridge University Press

figures cannot be relied upon. In such cases all data under suspicion must be treated alike. Only the prejudiced will keep the ones that please them and reject those that disturb. Sometimes data can be checked and that can be a great help but the check must be made on all those under suspicion, not just on ones that 'look wrong'.

Sometimes there is a group of data with a common mishap and sometimes it is possible to measure the severity of a defect discovered belatedly. If that is so, they can all be kept and adjusted at the analysis stage, perhaps by covariance, but again the adjustment must be used without favour.

It is not suggested that such decisions are usually made wrongly but only that they may be. There is danger in the method set out in Section 2 and it derives from the knowledge that there are statistical ways of testing aberrant data and of discarding any datum that fails the test. First, why was the datum thought aberrant? Because the experimenter did not like it or for some other reason? Also, in any large body of data there will be some residual that is larger than the rest. In any crowd there has to be someone who is farthest from the centre of attraction, but that person is still part of the crowd.

Also, as is pointed out above, the loss of data can often, though not necessarily, have serious effects. A datum should not be discarded lightly. If it is thought to be subject to only small disturbance it may be better to keep it.

There are, however, other ways of looking at the problem and that could lead to alternative approaches, always supposing that there is supporting evidence. For example, a difficulty arises if a trend is discovered across the experiment and shows the blocks to have been badly chosen. If a covariate can be constructed to allow for the trend, the situation can be retrieved, the blocks being retained or omitted in accordance with knowledge of any procedural changes that may have coincided with block boundaries.

Also, aberrant values can result from a bad choice of variate. In general, there is little conflict between measuring what should be measured and finding a quantity that satisfies the statistical criteria, but there are exceptions. There was once a mycologist who took some trouble to measure the area of lesions on apples from a spraying experiment. When the statistician inspected the data he wanted them transformed to their square roots to achieve Normality of distribution. Since the lesions were mostly approximately circular that transferred attention from the area to the radius, that is, the distance the infection had spread from the original airborne spore, a measure that has a clear biological meaning and is a reasonable indication of the efficacy of the treatments. The experimenter might just as well have measured each lesion by taking two diameters at right angles. It would have measured more effectively what he wanted to know, and it would have pleased the statistician. The situation is easier if there is long experience of the variate in question so that its distribution is predictable. Sometimes, as the present data show, a transformation is merely an expedient, and then opinions differ.

https://doi.org/10.1017/S0014479700003057 Published online by Cambridge University Press

REFERENCES

Bartlett, M. S. (1937). Some examples of statistical methods of research in agriculture and applied biology. *Journal of the Royal Statistical Society*, *Supplement*, 4:137–184.

Basford, K. E. & Tukey, J. W. (1995). Intercomparing residuals to find outliers in randomized blocks. *Australian Journal of Agricultural Research*, 46:451–461.

Burns, W. D., Charlton, P. J. & Li, W. K. (1978). Comparison of missing plot techniques. *Biometrical Journal*, 20:737–742.

Patterson, H. D. & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58:545–554.

Pearce, S. C. (1983). *The Agricultural Field Experiment*. Chichester: Wiley.

Pearce, S. C. (1992). Data analysis in agricultural experimentation. II. Some standard contrasts. *Experimental Agriculture*, 28:375–383.

Pearce, S. C. (1995). Some design problems in crop experimentation. I. The use of blocks. *Experimental Agriculture*, 31:191–203.

Rubin, D. B. (1972). A non-iterative algorithm for least squares estimation of missing values in any analysis of variance design. *Applied Statistics*, 21:136–141.

https://doi.org/10.1017/S0014479700003057 Published online by Cambridge University Press