A METHOD OF ESTIMATING THE YIELD OF A MISSING PLOT IN FIELD EXPERIMENTAL WORK.

By F. E. ALLAN, M.A. AND J. WISHART, D.Sc.

(*Statistical Department, Rothamsted Experimental Station, Harpenden.*)

It is essential in field experiments that the greatest care should be taken, so that reliance can be placed on the yield figures obtained, and normally there should be no case for rejection of any single plot. Yet it sometimes happens that the recorded yields are imperfect, owing to causes beyond the control of the experimenter. Thus a plot yield may be completely missing, or the plant may be so poor on a plot, owing to disease or for other reasons, that the recorded yield is not comparable with that of the other plots. Each such case should be carefully considered on its merits before a decision is reached as to whether the yield should be rejected or not. More than one lost plot would seriously jeopardise the accuracy of the experiment, and where modern methods of arrangement (Randomised Blocks and the Latin Square) are used, even one missing plot will disturb the symmetry and balance of the arrangement. In the former design, one plot of each treatment occurs in each block of land, and the variation between blocks as a whole is a measure of the differences in soil fertility over the experimental area. In the case of the Latin Square there are as many replicates as there are treatments, so arranged that no treatment occurs more than once in each row and column of the square.

Now when only one plot is missing something can be done in the way of estimating a value for it from the other plots, although it should be emphasised that there will be always some loss in accuracy as compared to an experiment with the full complement of plots. Cases that have occurred are: (*a*) a weight was missed from one plot of a Latin Square; (*b*) the corner plot of a similar square gave a very poor yield, the reason being that it adjoined the farm road and had evidently suffered badly from trampling. There was thus a good case for rejection of the particular plot; (*c*) in figures for a randomised block experiment recently submitted to this department, one figure, a protein determination, was missing and could not be retrieved.

Naturally we want to get as much information as possible from the experiment, even though it has suffered in the direction indicated. We could delete the whole block containing the missing plot, or the row or

column of the Latin Square, and this was actually done at a recent hay trial when it was found that one of the columns had suffered from the depredation of cows which had broken through the fence. Clearly this method sacrifices a good deal of information that the plots rejected would have been capable of furnishing, and what is needed is a means of utilising all the known plot values to form a best estimate of the missing yield. When only one plot is missing, such a method can be found by the application of the linear law which is the foundation of the analysis of variance procedure. For example in a randomised block experiment, if the yield of a particular plot is x, then the deviation of x from its mean $\bar{x}$ over the whole experiment is regarded as made up of a part due to the block in which it is situated, together with a part due to the treatment on that particular plot.

Thus $Y = b_p + t_q$, where Y is the supposed true deviation from the mean yield of the plot having treatment q in block p. For example, consider the following table of potato yields in lb.

	Treatments					
Block	1	2	3	4	Mean	
A.	139·0	219·0	200·5	145·0	175·88	
B.	197·5	205·0	206·0	182·5	197·75	
C.	156·0	229·5	210·0	245·5	210·25	
Mean	164·17	217·83	205·50	191·00	194·63	General mean

The linear law postulates that the yield of treatment 1 on Block A. is made up of mean yield of Block A. + mean of treatment 1 − general mean, *i.e.* 145·4. Actually it is 139·0, and the difference 6·4 represents the error. The sum of the squares of all such differences is, in fact, the sum of squares due to error in the analysis of variance. For the Latin Square there will be a part due to the row in which a particular plot is situated, a part for column, and a part for treatment, so that in this case the law is $Y = r_p + c_q + t_s$. Assuming such a linear law, we can now find an estimate for the missing yield.

(a) RANDOMISED BLOCKS.

The yield Y depends on the block factor b and the treatment factor t. Let $Y = b_p + t_q + k$ be the function, where k is constant throughout, b changes from block to block, and t from treatment to treatment, and let $b = 0$ in the block which contains the missing plot, and $t = 0$ for the treatment of which the missing one is a replicate, so that k is actually the estimated yield of the missing plot, and b and t represent deviations from it due to change in block and treatment in the other plots.

Suppose there are s treatments and n replicates of each. Then we have $ns - 1$ actual yields y, and we want to determine the constants so that the sum $S(y - Y)^2$ is a minimum taken over all the $ns - 1$ plots whose values are known. That is, we are to make $S(y - b_p - t_q - k)^2$ a minimum for variations of the constants b_p, t_q and k.

Minimising for variation in k gives

$$\overset{ns-1}{\underset{1}{S}}(y - b_p - t_q - k) = 0,$$

i.e.
$$\overset{ns-1}{\underset{1}{S}}(y) = sB + nT + (ns - 1)k = S, \text{ say,} \qquad \ldots\ldots(1)$$

where the summation extends over all the $ns - 1$ plots, and

$$B = b_1 + b_2 + \ldots + b_n, \quad T = t_1 + t_2 + \ldots + t_s,$$

it being remembered that B contains only $(n - 1)$ terms and T only $(s - 1)$ terms, one being missed from each since $b = 0$ for one block and $t = 0$ for one treatment, as was postulated above.

Minimising for variation in b_p, we get

$$\overset{s}{\underset{1}{S}}(y) = sb_1 + T + sk = y_1, \text{ say,}$$

where the summation extends over the s plots in the first block,

$$\overset{s}{\underset{1}{S}}(y) = sb_2 + T + sk = y_2 \text{ for summation over block 2,}$$

and so on for $(n - 1)$ blocks, omitting that in which $b = 0$. Adding these $(n - 1)$ equations, we get for the sum over $(n - 1)$ blocks

$$\overset{n-1}{\underset{1}{S}}(y_p) = sB + (n - 1)T + s(n - 1)k. \qquad \ldots\ldots(2)$$

Similarly, minimising for variations in t_q, we have, for the sum over $(s - 1)$ treatments,

$$\overset{s-1}{\underset{1}{S}}(y_q) = (s - 1)B + nT + (s - 1)nk, \qquad \ldots\ldots(3)$$

where y_q represents the sum of all the replicates of treatment t_q. The equations (1), (2) and (3) serve to determine B, T and k, and the contributory equations will give the values of b_p and t_q. It is sufficient for our purpose to find k. From the above equations we get

$$k(n - 1)(s - 1) = (n + s - 1)S - s.S_t - n.S_b, \qquad \ldots\ldots(A)$$

where

S_t = sum of treatment totals not including the treatment of which one plot is missing,

S_b = sum of block totals not including the block with the missing plot,

S = sum of all the known $(ns - 1)$ plot yields.

With the estimated value k for the missing plot we can proceed as usual with the analysis of variance. One degree of freedom has been lost from those for error due to fitting the value k, and the degrees of freedom now are: Blocks $n - 1$; Treatment $s - 1$; Error $(n - 1)(s - 1) - 1$; Total $ns - 2$.

Example. We can illustrate the process by the case referred to at the beginning, in which a figure was missing from a set of calculations of protein percentage in peas. There were nine different treatments and eight replicates.

Treatments

Block	1	2	3	4	5	6	7	8	9	Totals
A.	15·96	16·80	19·08	21·15	21·14	23·02	20·60	26·34	25·32	189·41
B.	21·62	24·44	26·33	[20·93]	20·22	20·87	19·21	20·04	21·30	[194·96]
C.	18·78	18·36	19·40	20·40	20·81	23·77	17·24	18·00	20·46	177·22
D.	20·18	20·52	20·42	19·96	20·69	20·90	21·68	21·90	20·68	186·93
E.	16·20	16·90	19·81	17·21	19·59	21·22	18·72	20·04	19·02	168·71
F.	18·64	20·33	22·71	19·29	18·33	19·06	20·04	21·31	22·19	181·90
G.	18·89	23·60	22·42	16·91	15·40	17·00	16·71	18·02	19·84	168·79
H.	19·22	20·47	20·98	18·06	20·34	21·38	16·83	15·66	17·21	170·15
Total	149·49	161·42	171·15	[153·91]	156·52	167·22	151·03	161·31	166·02	[1438·07]
Mean	18·686	20·178	21·394	19·239	19·565	20·903	18·879	20·164	20·753	19·973

The treatments in order from left to right are:

(1, 2, 3) No potash without superphosphate, and with single and double dressings of superphosphate.

(4, 5, 6) Single dressing of potash without superphosphate, and with single and double dressings of superphosphate.

(7, 8, 9) Double dressing of potash without superphosphate, and with single and double dressings of superphosphate.

The figures enclosed in brackets consist of the calculated value for the missing figure corresponding to treatment 4 on Block B., and the totals with this value included.

The equations for determining k are:

$$S = 1417 \cdot 14 = 71k + 9B + 8T, \qquad \text{from (1)}$$
$$S_b = 1243 \cdot 11 = 63k + 9B + 7T, \qquad \text{from (2)}$$
$$S_t = 1284 \cdot 16 = 64k + 8B + 8T, \qquad \text{from (3)}$$

and solving as in the general case, these give

$$k = \tfrac{1}{56} \{16 \times 1417 \cdot 14 - 9 \times 1284 \cdot 16 - 8 \times 1243 \cdot 11\}$$
$$= 20 \cdot 93.$$

Now put in this value of k for the missing figure in the table, and proceed

with the analysis of variance. There would in general be 71 degrees of freedom for a randomised block experiment with nine treatments and eight blocks, and the number here is therefore 70. The analysis of variance is now:

	Degrees of freedom	Sum of squares	Mean square	Standard error
Blocks	7	80·33	11·48	—
Treatments:				
Potash	2	0·48	0·24	—
Superphosphate ...	2	52·17	26·08	—
Interaction ...	4	3·93	0·98	—
Error	55	252·36	4·59	*0·437 or 2·19 %
Total	70	389·27	—	—

* Mean of 24 plots.

The experiment thus shows a significant effect of superphosphate in increasing the protein percentage of the peas, but no effect of potash, and the three superphosphate means (of 24 plots each) are 18·93, 19·97 and 21·02 respectively, with a standard error of 0·437.

(b) LATIN SQUARE.

The procedure for the Latin Square is similar, but we now write

$$Y = r_p + c_q + t_s + k$$

to show dependence on row, column and treatment. As before, k is the value of Y for the missing plot. Suppose there are n treatments, so that $n^2 - 1$ plot yields are known when one is missing.

$$S (y - r_p - c_q - t_s - k)^2$$

is now minimised, yielding the following equations.

For variations in k, summing over the whole area except the missing plot,

$$\overset{n^2-1}{\underset{1}{S}} (y) = (n^2 - 1) k + nC + nR + nT = S, \text{ say,} \quad \ldots\ldots(1)$$

where for convenience C has been written for the sum of all quantities like c_q, and so on for R and T. For columns, we have $(n - 1)$ equations of the type

$$\overset{n}{\underset{1}{S}} (y) = nk + nc_q + R + T = y_q, \text{ say,}$$

where the sum extends over column q. On addition these give

$$\overset{n-1}{\underset{1}{S}} (y_q) = (n - 1) nk + nC + (n - 1) R + (n - 1) T, \quad \ldots\ldots(2)$$

and similarly for the sum over $(n-1)$ rows

$$\overset{n-1}{\underset{1}{S}} (y_p) = (n-1)\,nk + (n-1)\,C + nR + (n-1)\,T, \quad ......(3)$$

and for treatments

$$\overset{n-1}{\underset{1}{S}} (y_s) = (n-1)\,nk + (n-1)\,C + (n-1)\,R + nT. \quad ......(4)$$

The equations (1), (2), (3) and (4) may now be solved for k, C, R and T. The solution in terms of row, column and treatment totals reduces to

$$k = \frac{1}{n-1}\,S_1 + \frac{2}{(n-1)\,(n-2)}\,S_2, \qquad ......(B)$$

where S_1 represents the sum of the yields of the $3\,(n-1)$ plots which agree with the missing plot in row, column or treatment, and S_2 is the sum of the plots which do not so agree.

　　An interesting illustration is given by a sugar beet experiment which was laid out in a 5×5 Latin Square. The yields are given in lb. for each plot of $\frac{1}{40}$ acre[1].

A.	B.	C.	D.	E.
306	556	369	332	396
B.	E.	D.	C.	A.
357	485	358	317	325
C.	D.	E.	A.	B.
309	467	367	275	413
D.	A.	B.	E.	C.
418	453	389	324	335
E.	C.	A.	B.	D.
503	572	346	397	[279]

　　The yield in the bottom right-hand corner is markedly low, especially when we consider the relatively high yields for treatment D on the other plots; and actually a note was appended with the figures commenting on the very poor plant on this particular plot and saying that it was due to trampling during the growing season. Accepting the figure of 279 the analysis of variance was as follows:

	Degrees of freedom	Sums of squares	Mean square	$\frac{1}{2}\log_e \times$ Mean square	Standard error
Rows	4	9,279·04	2,319·76	—	—
Columns	4	97,905·44	24,476·36	—	—
Treatment	4	22,266·64	5,566·66	4·3123	—
Error	12	28,384·72	2,365·39	3·8843	*21·75 or 5·64 %
Total	24	157,835·84	6,576·49	—	—

* Mean of 5 plots.

[1] See Wishart and Hines, "Fertiliser Trials on the Ordinary Farm," *J. Min. Agric.*, Sept. 1929, pp. 524–32.

If this was all that could be done, we should be forced to conclude that the differences between the treatment means are not significant in the light of their standard error. Let us, however, discard the misleading value 279 and replace it by a value estimated on the assumption of the linear law above. We find for the estimated value,

$$k = \tfrac{1}{4}.(4862) - \tfrac{1}{6}.(4507)$$
$$= 464 \cdot 3.$$

Take this as 464, since all the weights are only accurate to the nearest pound. One degree of freedom is lost, and the analysis of variance becomes

	Degrees of freedom	Sums of squares	Mean square	$\tfrac{1}{2}\log_e \times$ Mean square	Standard error
Rows	4	27,142·64	—	—	—
Columns	4	89,943·04	—	—	—
Treatment	4	22,148·24	5,537·06	4·3096	—
Error	11	11,897·52	1,081·59	3·4931	*14·71 or 3·74 %
Total	23	151,131·44	—	—	—

* Mean of 5 plots.

The part due to error is now so much smaller that, despite the fact that the number of degrees of freedom is reduced, the z test shows that there is a significant response to treatment[1]. The value of z is 0·8165, and since a value of z greater than 0·6055 occurs by chance in only 5 per cent. of cases, this value is certainly significant. Thus it has been possible to gain positive results from an experiment which at first appeared to have been spoilt by the unreliable figure for the yield of one plot. The results are:

Average yield	No nitrogen	Sulphate of ammonia	Muriate of ammonia	Cyanamide alone	Cyanamide and nitrate of soda	Mean	Standard error
Tons per acre	6·1	7·4	7·3	6·8	7·5	7·0	0·26
Per cent.	86·7	105·5	103·7	96·7	107·4	100·0	3·74

The other possible procedure, of discarding the whole of the last column, in which the damaged plot occurred, and treating the remainder as a randomised block experiment, gives, like the first analysis, a standard error so high that nothing could be inferred from the experiment. We could not, indeed, expect much from such an experiment with only four replicates. This last procedure is moreover open to the objection that the four columns used are not in fact randomised blocks, but are subject to the further restriction necessary to make them part of a Latin Square.

[1] See Wishart and Hines, *loc. cit.*

Summary.

In cases of field experiments when one plot is missing, a method has been developed for furnishing an estimate of the yield of the missing plot, based on all the other values. The calculation is given for (*a*) a Randomised Block experiment, and (*b*) a Latin Square arrangement. In both cases the actual arithmetic is very simple.

The steps in the procedure are:

(1) Determine the desired value by an application of the equation (*A*) or (*B*) according as the experiment was arranged in Randomised Blocks or in a Latin Square.

(2) Proceed as usual with the analysis of variance, using the estimated figure for the missing yield, and remembering to deduct one from the number of degrees of freedom ascribable to error.