THE ESTIMATION OF MISSING OBSERVATIONS IN SEVERAL EXPERIMENTAL DESIGN

by

CHIA-SHENG WANG

B. B. A., National Taiwan University, 1962

A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1967

Approved by:

Young A. Koh

Major Professor

TABLE OF CONTENTS

INTRODUCTION . . . . (1)

NOTATION . . . . (2)

GENERAL SOLUTION . . . . (3)

RANDOMIZED BLOCK DESIGN . . . . (5)

RANDOMIZED BLOCK DESIGN WITH REPLICATION WITHIN UNITS . . . . . (10)

CROSS-OVER DESIGN . . . . (15)

LATIN SQUARE DESIGN . . . . (16)

SPLIT PLOT DESIGN . . . . (16)

DISCUSSION . . . . (21)

INTRODUCTION

The simple analysis of most experimental designs depends upon their fully balanced arrangement. But in experimental work it frequently happens that one or more experimental units is missing from the data, or has to be rejected because of conditions outside the control of the experimenter. It should be cautioned that observations should be rejected in the analysis of results only under extreme circumstances, when it is quite obvious that the treatment being studied is not responsible for the apparently anomalous results (Anderson, 1946).

One of the first papers on the subject of estimating the yield of a missing unit in field experimental work was published by Allan and Wishart (1930). They derived formulas and illustrated their use for a single missing plot in a randomized block and in a latin square experiment. These methods were extended by Yates (1933) to cover several missing units in a given experiment.

The formula given by Yates for estimating the yield of a single missing unit in a randomized block experiment is

$$Y = \frac{(rB + tT - G)}{(r - 1)(t - 1)}$$

where r = the number of blocks,

t = the number of treatments,

B = the total yield of the remaining units in the block where the missing unit appears,

T = the total of the yields of the treatments with the missing unit,

G = the grand total,

Similarly for a single missing unit in a latin square

$$y = \frac{r(R + C + T) - 2G}{(r - 1)(r - 2)}$$

where r = the number of rows, columns and treatments,

R = the total number of yields of the remaining units in the row in which the missing unit appears,

C = the total number of yields of the remaining units in the column in which the missing unit appears.

In this report the problem of calculating these missing values will be investigated from a general point of view for the following designs:

(1) randomized block,

(2) randomized block design with replication within units,

(3) cross-over,

(4) latin square,

(5) split-plot.

NOTATION

Vectors - lower case letters in Clarendon type.

Matrices - capital letters in Clarendon type.

Transposition of matrices and vectors is indicated by a prime.

Summation - two summation symbols will be used:

(1) $\Sigma\, x_i$ where $i = 1, 2, \ldots, n$

(2) $\Sigma_{(i)}\, x_i$ to mean summation over some subset of the subscript values.

Associated observations - an observation in a design is denoted by $x_{ijk\ldots}$, where the number of subscripts indicates the order of classification involved. Thus an observation in a randomized block requires two subscripts for its specification, and an observation in a Graeco-latin square requires four.

If one consider two observations they may have certain subscripts in common. This property enables one to describe the relation between pairs of observations; observations related in a specific way will be called associates. For example, one consider two observations x_{fgh} and x_{uvw}. The following classes of associates can be recognized:

(1) Zero-order associates: $f \neq u$, $g \neq v$, $h \neq w$;

(2) 1st-order associates: (i) i associates $f=u$, $g \neq v$, $h \neq w$;
(ii) j associates $f \neq u$, $g=v$, $h \neq w$;
(iii) k associates $f \neq u$, $g \neq v$, $h=w$;

(3) 2nd-order associates: (i) ij associates $f=u$, $g=v$, $h \neq w$;
(ii) ik associates $f=u$, $g \neq v$, $h=w$;
(iii) jk associates $f \neq u$, $g=v$, $h=w$.

Particular associates may or may not exist depending on the experimental design.

A potential observation in a design will be indicated by a capital letter $X_{ij\ \ldots}$, if this is in fact a missing observation it will be indicated in lower case $x_{ij\ \ldots}$.

GENERAL SOLUTION

Let x_i $(i = 1, 2, \ldots, p)$ be p independent missing values. Then the error sum of squares is a quadratic function of these values, $F(x_1, x_2, \ldots, x_p)$, and the vector of the partial derivative of this function is given by Biggers (1959), that is:

$$2(x'A - q')/N. \tag{1}$$

where x is a column vector of the p missing values,

A is a p x p symmetric matrix whose elements are determined by the experimental design and the distribution of the missing values,

q is a column vector whose elements are calculated from the available observations, N is a constant determined by the experimental design.

The values of the x_i, where x_i is a minimum, are given by the solution of p simultaneous equations obtained by equating each of its partial derivatives to zero. These are, in matrix notation:

$$Ax = q. \tag{2}$$

Several methods may be used to solve equation (2). Most textbooks on the design and analysis of experiments give the formula for one missing observation in randomized block, cross-over and latin-square designs and the special formula for two or three missing observations have also been published for the randomized block design (Baten, 1939, 1952; Federer, 1951, 1955) and the latin-square design (Federer, 1955). Usually, where more than one missing observation exist in a design, it is recommended that the estimates be obtained by an iterative process based on the formula for one missing observation. None of these methods gives the explicit determination of the simultaneous equations. In this report rules will be given which enable equation (2) to be written down rapidly for the different experimental designs. The main advantage of this procedure is that it enables the nature of the equations to be examined and this leads to simplifications in several important complex cases.

The matrix method has been the subject of a paper by Thompson (1956) on missing values in the randomized block designs. This paper, however, only dealt with a restricted distribution of missing values, presumably because the values of A^{-1} are written down instead of A. Wilkinson (1957) has shown how the matrix A can be used to obtain exact variances of treatment comparisons.

RANDOMIZED BLOCK DESIGN

Suppose an experiment consists of r blocks each containing t plots to which t treatments are allotted at random. Let X_{ij} ($i = 1, 2, \ldots, r$; $j = 1, 2, \ldots, t$) be the observation which is in the ith block and receiving the jth treatment. At the end of the experiment let p of these values be missing. It should be noted that the missing observations are indicated by only some of the values of the subscripts.

Let the ith block and the jth treatment contain c_i and d_j missing observations respectively. Then

$$p = \Sigma\, c_i = \Sigma\, d_j.$$

Missing observations in the same ith block are called i associates or block associates and missing observations receiving the same jth treatment are called j associates or treatment associates. The missing observation x_{gh} will have therefore $(c_g - 1)$ block associates, $(d_h - 1)$ treatment associates and $p - (c_g + d_h - 1)$ zero associates.

Let B_i = total of available observations in the ith block,

T_j = total of the available observations receiving the jth treatment,

G = total of all available observations.

The error term in the analysis of variance is given by

$$\Sigma_{(i)}\Sigma_{(j)}x_{ij}^2 - \Sigma_{(i)}(B_i + \Sigma_{(j)}x_{ij})^2/\,t - \Sigma_{(j)}(T_j + \Sigma_{(i)}x_{ij})^2/\,r$$
$$+ (G + \Sigma_{(i)}\Sigma_{(j)}x_{ij})^2/\,rt + \text{constant}. \qquad (3)$$

Partial differentiation of (3) with respect to a particular missing observation, x_{gh}, and equating the derivative to zero, gives the equation

$$rtx_{gh} - r\,\Sigma_{(j)}x_{gj} - t\,\Sigma_{(i)}x_{ih} + \Sigma_{(i)}\Sigma_{(j)}x_{ij} = rB_g + tT_h - G.$$

This function can be rewritten as follow:

$$rtx_{gh} - r(x_{gh} + \sum_{(j) \atop j \neq h} x_{gj}) - t(x_{gh} + \sum_{(i) \atop i \neq g} x_{ih}) + \sum_{(i) \atop i \neq g} \sum_{(j) \atop j \neq h} x_{ij} + \sum_{(i) \atop i \neq g} x_{ih}$$

$$+ \sum_{(j) \atop j \neq h} x_{gj} = rB_g + tT_h - G$$

Separation of terms corresponding to zero, i and j associates yields

$$(r - 1)(t - 1)x_{gh} + \overbrace{(1 - r)\sum_{(j) \atop j \neq h} x_{gj}}^{\text{i associates}} + \overbrace{(1 - t)\sum_{(i) \atop i \neq g} x_{ih}}^{\text{j associates}} + \overbrace{\sum_{(i) \atop i \neq g} \sum_{(j) \atop j \neq h} x_{ij}}^{\text{zero associates}}$$

$$= rB_g + tT_h - G. \qquad (4)$$

Expression (4) enables one to write down the coefficients of the x_{ij}, the B_g and T_h, and hence the matrix A and the vector q in (2).

The elements of matrix A are given by:

(i) $(r - 1)(t - 1)$ for the missing observation under consideration,

(ii) $(1 - r)$ for the block associates,

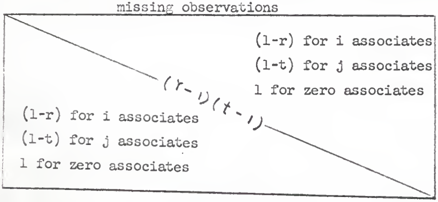
(iii) $(1 - t)$ for the treatment associates,

(iv) 1 for the zero associates.

Usually, in practical situations the matrix A will be nonsingular, and thus the equation (2) will have a unique solution.

The general form of matrix A for the randomized block design will be like Figure 1.

Figure 1. The general form of matrix A for randomized block design.



Suppose there are five missing observations, they are: x_{12}, x_{13}, x_{35}, x_{45}, and x_{67}. Then one can easily find the matrix A as follows:

	x_{12}	x_{13}	x_{35}	x_{45}	x_{67}
x_{12}	$(r-1)(t-1)$	$(1-r)$	1	1	1
x_{13}	$(1-r)$	$(r-1)(t-1)$	1	1	1
x_{35}	1	1	$(r-1)(t-1)$	$(1-t)$	1
x_{45}	1	1	$(1-t)$	$(r-1)(t-1)$	1
x_{67}	1	1	1	1	$(r-1)(t-1)$

Example 1. Table 1 shows the results of a randomized block experiment ($r = 8$; $t = 5$). The missing values are calculated as follows:

(1) Construct a table so the rows correspond to separate blocks. This is essential to avoid confusion in determining the associates.

(2) Obtain the coefficients of the associates.

(3) Insert in the table of results the letter x_{ij} for the missing values with subscripts determined by the position of the missing value in the table, e.g. x_{35} is the missing value in the third block receiving the fifth treatment. Space should be left to insert the estimated missing value.

(4) Obtain the B_i, T_j and G from the available observations and write at the border of the table,

(5) Prepare a table as in Table 2. List the subscripts in a row at the top and a column at the left-hand side of the table. By comparing the subscripts in pairs, determine which subscripts are common. This

determines the type of association, and the appropriate coefficient is written down at the intersection of the corresponding row and column. At the right of the table 2 columns for the B_i, T_j and G are prepared. The i subscript in the left-hand column determines the B_i total in a given row, while the j subscript determines the T_j total. Beneath the B_i and T_j columns the values of r and t are written down respectively, and beneath the G column the value -1 is written.

Table 1. The wet weight in mg (x) of embryonic chick tibiae after cultivation on a chemically defined medium containing five different concentrations of glucose. Treatment numbers are placed in parentheses. The variate analysed is log 10x (Biggers,1959)

Block	Glucose concentration (mg./ml.)					B	Completed block totals
	0.5(1)	1.0(2)	2.0(3)	4.0(4)	8.0(5)		
I	0.88	1.15	1.33	1.39	1.57	-	6.32
II	1.06	1.22	1.63	1.54	1.37	-	6.82
III	0.97	1.34	(x=1.50)	1.66	(x=1.53)	3.97(B_3)	7.00
IV	1.09	1.21	1.16	1.50	1.48	-	6.44
V	1.14	1.37	1.58	1.52	1.44	-	7.05
VI	1.13	1.33	1.65	1.57	1.47	-	7.15
VII	1.00	1.21	(x=1.44)	1.45	1.52	5.18(B_7)	6.62
VIII	1.12	1.30	1.35	(x=1.56)	1.61	5.38(B_8)	6.94
T	-	-	8.70(T_3)	10.63(T_4)	10.46(T_5)	48.31(G)	-
Completed treatment totals	8.39	10.13	11.64	12.19	11.99	-	54.34

Table 2. Calculation of A and q

Subscripts	A				q		
	33	35	73	84	B	T	G
33	28	-7	-4	1	3.97	8.70	48.31
35	-7	28	1	1	3.97	10.46	48.31
73	-4	1	28	1	5.18	8.70	48.31
84	1	1	1	28	5.38	10.63	48.31
					r=8	t=5	-1

(6) Invert the matrix A,

(7) Write down the solution as follows:

$$\begin{bmatrix} x_{33} \\ x_{35} \\ x_{73} \\ x_{84} \end{bmatrix} = \overset{A^{-1}}{\begin{bmatrix} 0.039 & 0.010 & 0.005 & -0.002 \\ 0.010 & 0.038 & 0.000 & -0.002 \\ 0.005 & 0.000 & 0.037 & -0.001 \\ -0.002 & -0.002 & -0.001 & 0.036 \end{bmatrix}} \overset{q}{\begin{bmatrix} 26.95 \\ 35.75 \\ 36.63 \\ 47.88 \end{bmatrix}}$$

q is obtained by row by row multiplication of each row of the right-hand side of Table 2 with the bottom row. Thus, in the example, the upper entry of q is given by

$$8 \times 3.97 + 5 \times 8.70 - 48.31 = 26.95$$

The values of the x_{ij} are then obtained by standard row by column multiplication. In the example $x_{33} = 1.50$; $x_{35} = 1.53$; $x_{73} = 1.44$; $x_{84} = 1.56$.

(8) The values of the x_{ij} are inserted in Table 1, and the block and treatment totals completed. The standard analysis of variance then follows. In this example 4 D.F. are subtracted from the error term.

Special case: The process of calculating A^{-1} can be simplified if A is a p x p matrix of the form:

$$\begin{bmatrix} a & b & . & . & b \\ b & a & . & . & b \\ . & . & . & . & . \\ . & . & . & . & . \\ b & b & . & . & a \end{bmatrix}. \tag{5}$$

This is a circulant matrix and hence the sum of the elements of a row is a factor of its determinant. Thus if $a + b(p - 1) = 0$ the matrix is singular, and A^{-1} will not be found.

The inverse of the matrix A is given by

$$A^{-1} = \frac{1}{(a-b)[a+b(p-1)]} \begin{bmatrix} a + b(p - 2) & -b & . & . & -b \\ -b & a + b(p - 2) & . & . & -b \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -b & -b & & . & a + b(p - 2) \end{bmatrix}.$$

RANDOMIZED BLOCK DESIGN WITH REPLICATION WITHIN UNITS

Suppose on has r blocks each containing t plots to which t treatments are allotted at random, and in each cell there are s independent observations. The analysis of variance takes the form shown in Table 3. At the present time there is considerable discussion on the linear models which are assumed in the analysis of this design (Anderson and Bancroft, 1952; Wilk and Kempthron, 1955). The models determine the expectations of the mean squares in the analysis of Table 3. There are three error terms to consider:

(1) The within-plots sum of squares (E_1) (case 1),

(2) The blocks x treatments sum of squares (E_2) (case 2),

(3) ($E_1 + E_2$) (case 3).

Table 3. Analysis of variance for a r x t randomized block design with s observations per plot

Source of variation	D.F.	S.S.
Between blocks	(r-1)	(B-C)
Between treatments	(t-1)	(T-C)
Blocks x treatments interaction(E_2)	(r-1)(t-1)	(P-B-T+C)
Within plots (E_1)	rt(s-1)	(S-P)
Total	(rst-1)	(S-C)

Where S = total sum of squares, C = correction for the mean, B = block sum of squares, T = treatment sum of squares, P = plot sum of squares.

Let X_{ijk} (i = 1, 2, ..., r; j = 1, 2, ..., t; k = 1, 2, ..., s) be the kth observation on the jth treatment in the ith block, and let p_{ij} be the number of missing observations within each plot. Then the total number of missing observations is $p = \sum p_{ij}$. The following associates exist: zero, i, j and ij.

Let P_{ij} = total values of the available observations in the ijth plot,

B_i = total values of the available observations in the ith block,

T_j = total values of the available observations receiving the jth treatment,

G = total values of all available observations.

The three sets of equations for estimating the missing values when each of the error terms are minimized, have been obtained by an analysis similar to that applied to the randomized block design. The rules which determine the elements of A and q for each case are shown in Table 4.

If one orders the elements of A in groups corresponding to the ij subscript, one can rewrite (2) in general:

$$\begin{bmatrix} A^{11}_{11} & A_{12} & \cdot\ \cdot & A_{1.rt} \\ A_{12} & A^{12}_{22} & \cdot\ \cdot & A_{2.rt} \\ \cdot & \cdot & \cdot\ \cdot & \cdot \\ \cdot & \cdot & \cdot\ \cdot & \cdot \\ A_{1.rt} & A_{2.rt} & \cdot\ \cdot & A^{rt}_{rt.rt} \end{bmatrix} \cdot \begin{bmatrix} x^{11} \\ x^{12} \\ \cdot \\ \cdot \\ x^{rt} \end{bmatrix} = \begin{bmatrix} q^{11} \\ q^{12} \\ \cdot \\ \cdot \\ q^{rt} \end{bmatrix} \tag{6}$$

Where $A^{ij}_{\alpha B}$ ($\alpha = B$) are submatrices along the main diagonal, one corresponding to each plot. Each is a square symmetric matrix of order p_{ij}. The superscript denotes the plot. While the subscript denotes the position of the submatrix,

$A_{\alpha B}$ ($\alpha \neq B$) are rectangular submatrices,

x^{ij} is a vector of the p_{ij} missing values in plot ij,

q^{ij} is a vector of the p_{ij} elements calculated from the available observations.

Table 4. Coefficients of associates and elements of q for the three cases in the randomized block design with replication within units.

Associate	Coefficient		
	Case 1 (E_1)	Case 2 (E_2)	Case 3 ($E_1 + E_2$)
Missing value	(s-1)	(r-1)(t-1)	(rst-r-t+1)
ij	-1	(r-1)(t-1)	(1-r-t)
i	0	(1-r)	(1-r)
j	0	(1-t)	(1-t)
zero	0	1	1

Elements of q

Case 1	P_{fg}
Case 2	$rB_f + tT_g - rtP_{fg} - G$
Case 3	$rB_f + tT_g - G$

If a plot has no missing values, the submatrices of the corresponding rows and columns in (6) vanish. The elements of the submatrices along the main diagonal are determined by the coefficients for the missing value and its ij associates.

We now consider the three cases discussed above.

Case 1. The matrix of (6) is of the form

$$\begin{bmatrix} A_{11} & & & & Z \\ & A_{12} & & & \\ & & \cdot & & \\ & Z & & \cdot & \\ & & & & A_{rt\cdot rt} \end{bmatrix} \tag{7}$$

Since the inverse of (7) is

$$\begin{bmatrix} A_{11}^{-1} & & & & \\ & A_{12}^{-1} & & Z & \\ & & \cdot & & \\ & Z & & \cdot & \\ & & & & A_{rt\cdot rt}^{-1} \end{bmatrix} \tag{8}$$

which means that the estimates of the missing values in the ijth plot is independent of the missing values in all other plots. Therefore the missing values for the ijth plot are given by the solution of

$$\underset{p_{ij} \times p_{ij}}{\begin{bmatrix} (s-1) & -1 & \cdot & \cdot & -1 \\ -1 & (s-1) & \cdot & \cdot & -1 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ -1 & -1 & \cdot & \cdot & (s-1) \end{bmatrix}} \begin{bmatrix} x_{ij1} \\ x_{ij2} \\ \cdot \\ \cdot \\ x_{ijp_{ij}} \end{bmatrix} = \begin{bmatrix} P_{ij} \\ P_{ij} \\ \cdot \\ \cdot \\ P_{ij} \end{bmatrix} . \tag{9}$$

The inverse of the matrix A can be written in general form:

$$\begin{bmatrix} \frac{s-(p_{ij}-1)}{s(s-p_{ij})} & \frac{1}{s(s-p_{ij})} & \cdots & \frac{1}{s(s-p_{ij})} \\ \frac{1}{s(s-p_{ij})} & \frac{s-(p_{ij}-1)}{s(s-p_{ij})} & \cdots & \frac{1}{s(s-p_{ij})} \\ \cdot\ \cdot & \cdot\ \cdot & \cdots & \cdot\ \cdot \\ \frac{1}{s(s-p_{ij})} & \frac{s-(p_{ij}-1)}{s(s-p_{ij})} & \cdots & \frac{s-(p_{ij}-1)}{s(s-p_{ij})} \end{bmatrix} . \tag{10}$$

This has the solution $p_{ij} / (s - p_{ij})$ for all the x_{ijk}. Thus in this case the missing observations are estimated by the average of the available observations in the plot. If all the values in a plot are missing $p_{ij} = s$ and the equations have no straight-forward solution (Biggers, 1959).

When $r = 1$, one can have the case of fully randomized design. Thus an estimate of a missing values, if required, is the average of the results available on the particular treatment.

Case 2. The submatrices, in this case, have equal elements. The $A^{ij}_{\alpha B}$ ($\alpha = B$) have elements equal to $(r-1)(t-1)$ while the $A_{\alpha B}$ ($\alpha \neq B$) have elements equal to $(1-r)$, $(1-t)$ or 1, according to i, j or zero associates. Thus, to solve (6), one may allot within each plot $(p_{ij} - 1)$ arbitrary values and then solve for the remainder. In practice, therefore, one would insert arbitrary values for all but one missing value in each plot, say $p_{ij} / (s - p_{ij})$ for $s \neq p_{ij}$, and then write out (6) as though one value is missing in each plot. Since, in case 1, the missing observations within plot are estimated by the average of the available observations in the plot. Alternatively, if one works entirely with plot totals, one can use the formula for a randomized block design to estimate a missing plot ignoring all the available observations on the plot.

Case 3. Here the matrix of (6) is assumed to be non-singular even if all observations are missing from a plot, thus allowing estimates of these values to be made.

CROSS-OVER DESIGN

Consider an r x t design, where each of t columns is a separate replicate of the r treatments, subject to the condition that each treatment occurs an equal number of times (a) in each row, then $t = ar$

Denote each observation by $X_{ijk}(i,\ k = 1,\ 2,\ \ldots,\ r;\ j = 1,\ 2,\ \ldots,\ t)$. Let p observations be missing. In this design, the following associates exist: zero, i, j, k, ik.

Let R_i = total of the available observations in the ith row,

C_j = total of available observations on the jth column,

T_k = total of the available observations receiving the kth treatment,

G = total of all available observations.

The elements of matrix A are given by:

(i) $(r-1)(t-2)$ for the missing observation under consideration,

(ii) $2(1-r)$ for the ik associates,

(iii) $(2-r)$ for the i and k associates,

(iv) $(2-t)$ for the j associates,

(v) 2 for the zero associates.

The element of q is given by

$$rR_f + tC_g + rT_h - 2G.$$

When $r = 2$ the coefficient for i and k associates is zero. This leads to independence of the equations for estimating the missing values in some instances, thus allowing the formula for a single missing value to be used separately for each plot.

LATIN SQUARE DESIGN

The formula for a t x t latin-square design can be obtained from those for the cross-over design by putting a = 1. Under this restriction ik associates do not exist.

The elements of matrix A are given by:

(i) (t-1)(r-2) for the missing observation under consideration,

(ii) (2-t) for the i, j and k associates,

(iii) 2 for the zero associates.

The element of q is then

$$t(R_f + C_g + T_h) - 2G .$$

SPLIT PLOT DESIGN

Whole-plot treatments arranged in randomized blocks

In field experiments an extra factor is sometimes introduced into an experiment by dividing each plot into a number of parts.

Now, let us consider a design with r blocks containing t plots to which t whole-plot treatments are allotted at random. Further let s split-plot treatments be allotted at random within each whole plot. For example, if the experiment is planned originally to test a factor A with five levels, and divided each plot into halves permits an extra factor B at two levels. The plan (after randomization) might appear as shown as follows:

Block 1

a_3	a_2	a_0	a_1	a_4
b_0	b_1	b_1	b_1	b_0
b_1	b_0	b_0	b_0	b_1

Block 2

a_1	a_4	a_0	a_3	a_2
b_0	b_0	b_0	b_1	b_1
b_1	b_1	b_1	b_0	b_0

Block 3

a_4	a_3	a_0	a_1	a_2
b_0	b_1	b_1	b_0	b_0
b_1	b_0	b_0	b_1	b_1

The analysis of variance of this design is given in Table 5. There are two error terms to be considered: (1) E_1 corresponding to the whole-plots (case 1), and (2) E_2 corresponding to the split-plot (case 2).

Table 5. Analysis of variance for a split-plot design with whole-plot treatments arranged in randomized blocks

Source of variation	D.F.	Sum of squares
Between blocks	(r-1)	(B-C)
Between whole-plot treatments (A)	(t-1)	(S-C)
Error (E_1)	(r-1)(t-1)	(W-B-S+C)
Total	(rt-1)	(W-C)
Between split-plot treatments (B)	(s-1)	(T-C)
AB interaction	(s-1)(t-1)	(I-S-T+C)
Error (E_2)	t(s-1)(r-1)	(Z-W-I+S)
Total	(rst-1)	(Z-C)

Where Z = total sum of squares of split-plots, W = total sum of squares, C = correction for the mean, B = block sum of squares, S = whole-plot treatment sum of squares, T = split-plot treatment sum of squares, I = treatment combination sum of squares.

Let one denote the observation made on the kth split-plot treatment in the ijth plot by X_{ijk} $(i = 1, 2, \ldots, r;\ j = 1, 2, \ldots, t;\ k = 1, 2, \ldots, s)$.

At the end of the experiment let p split-plots be missing.

The following associates exist: zero, i, j, k, ij, ik, jk.

Case 1. Table 5 shows that the expression for the error term is the same as that in case 2 of the generalized randomized block design. The solution in the two cases is identical. One ignores the available observations and estimate new whole-plot values using the formula for a randomized block design.

Case 2. Let W_{ij} = total of the available observations in the ijth plot,

I_{jk} = total of the available observations receiving the jth whole-plot treatment,

S_j = total of the available observations receiving the jth whole-plot treatment.

The elements of matrix A are given by:

(i) (r-1)(s-1) for the missing observation under consideration,

(ii) (1-s) for the jk associates,

(iii) (1-r) for the ij associates,

(iv) 1 for the j associates,

(v) 0 for the zero, i, k and ik associates.

The element of q is given by

$$rW_{fg} + sI_{gh} - S_g.$$

Example 2. Table 6 shows the results of a split-plot experiment where the whole-plot treatments are arranged in randomized blocks ($r = 4$; $t = 5$; $s = 2$). Missing values have been artificially chosen to provide the example, Table 7 is prepared from the data of Table 6 in exactly the same way as discribed in Example 1, except that the subscripts have been ordered using the middle (j) subscript. From the matrix A it is seen that the estimation of x_{121} and x_{222} is independent of the estimation of x_{132}

and x_{232}. Thus, instead of inverting a 4 x 4 matrix, one may invert two 2 x 2 matrices considerably simplifying the problem. The estimates are given by

$$\begin{bmatrix} x_{121} \\ x_{222} \end{bmatrix} = 1/8 \begin{bmatrix} 3 & -1 \\ -1 & 3 \end{bmatrix} \begin{bmatrix} 4.74 \\ 4.74 \end{bmatrix}$$

$$\begin{bmatrix} x_{132} \\ x_{232} \end{bmatrix} = 1/8 \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} 2.56 \\ 3.76 \end{bmatrix} .$$

Thus $x_{121} = 1.19$; $x_{222} = 1.19$; $x_{132} = 1.43$; $x_{232} = 1.73$.

Table 6. The wet weight in mg(x) of embryonic chick titiae after cultivation on a chemically defined medium containing different concentrations of glucose or mannose. Treatment numbers are placed in parentheses. The variate analysed is log 10x (Biggers, 1959).

Block	Split-plot treatment	Whole-plot treatment Hexose concentration (mg./ml.)				
		0.5(1)	1.0(2)	2.0(3)	4.0(4)	8.0(5)
I	Glucose(1)	0.88	(x_{121}=1.19)	1.33	1.39	1.57
	Mannose(2)	0.78	1.15	(x_{132}=1.43)	1.42	1.51
II	Glucose(1)	1.06	1.22	1.63	1.54	1.37
	Mannose(2)	1.09	(x_{222}=1.19)	(x_{232}=1.73)	1.47	1.44
III	Glucose(1)	0.97	1.34	1.44	1.66	1.59
	Mannose(2)	0.90	1.16	1.44	1.48	1.59
IV	Glucose(1)	1.09	1.21	1.16	1.50	1.48
	Mannose(2)	1.07	1.32	1.36	1.41	1.41

Table 7. Calculation of A and q

Subscripts	A				q		
	121	222	132	232	W_{ij}	I_{ik}	S_i
121	3	1	0	0	1.15	3.77	7.40
222	1	3	0	0	1.22	3.63	7.40
132	0	0	3	-1	1.33	2.80	8.36
232	0	0	-1	3	1.63	2.80	8.36
					r=4	s=2	-1

DISCUSSION

In general , the missing observations in experimental data are estimated by mininizing the Error Sum of squares (Biggers, 1959). When a value is missing, one can use the formula derived by Yates in 1933. When more than one observations are missing, one can use the equation of the form Ax = q. Since the Error Sum function, $F(x_1, x_2, \ldots, x_p)$, is a quadratic form, then when the partial derivatives are set equal to zero, the following equation is obtained:

$$Ax = q .$$

If A is non-singular, then x is given directly by

$$x = A^{-1}q.$$

When several values are missing, the A matrix becomes quite large but with the aid of high-speed computers the inverse can easily be found.

Kendall (1946) has pointed out that using the estimated values obtained by this general method for the analysis of variance test does not affect the test of significance to any serious degree.

Yates (1933) has shown that the estimates of treatment and block effects are exactly the same as those obtained by the correct least squares procedure, and the Error Sum of squares is exactly the same as given by the correct procedure.

ACKNOWLEDGMENT

The writer wishes to express his gratitude to his major professor, Dr. Young O. Koh, for all his advice and excellent criticism that had helped to improve this report.

The writer would also like to express his appreciation to Dr. A. M. Feyerherm for his valuable comments and criticisms on this report.

REFERENCES

Allan, F. E. & Wishart, J. (1930). J. Agric. Sci. 20, 399.

Anderson, R. L. (1946). Biometrics, 2, 41.

Anderson, R. L. & Bancroft, T. A. (1952). Statistical Theory in Research. London: McGraw-Hill.

Bartlett, M.S. (1937). J. R. Statist. Soc. Suppl. 4, 137.

Baten, W. D. (1952). Biometrics, S, 42.

Biggers, J. D. (1959). Biometrika, 46, 91.

Coons, I. (1957). Biometrics. 13,387.

Federer, W. T. (1955). Experimental Design. New York: Macmillan and Co.

Fox, L. (1950). J. R. Statist. Soc. B, 12, 120.

Fox, L. & Hayes, J. G. (1951). J. R. Statist. Soc. B, 13, 83.

Healy, M. & Westmacott, M. (1956). Appl. Statist. 5, 203.

Kendall, M.A. (1946). The Advanced Theory of Statistics. 2, 229.

Khargonkar, S. A. (1948). J. Indian Soc. Agric. Statist. 1, 147.

Thompson, H. R. (1956). Biometrics, 12, 241.

Wilk, M. B. & Kempthorne, O. (1955). J. Amer. Statist. Ass. 50, 1144.

Wilkinson, G. N. (1957). Biometrics, 13, 363.

Wilkinson, G. N. (1958a). Biometrics, 14, 257.

Wilkinson, G. N. (1958b). Biometrics, 14, 360.

Yates, F. (1933). Exp. J. Exp. Agric. 1, 129.

THE ESTIMATION OF MISSING OBSERVATIONS
IN SEVERAL EXPERIMENTAL DESIGN

by

CHIA-SHENG WANG

B. B. A., National Taiwan University, 1962

AN ABSTRACT OF A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1967

Approved by:

Young C. Koh

Major Professor

In experimental work it frequently happens that one or more experimental units is missing from the data. One of the first papers on the subject of estimating a missing unit in field experimental work was published by Allan and Wishart (1930), and had been extended by Yates (1933) to cover several missing units in a given experiment. The values which were inserted to minimize the error sum of squares.

The missing values were calculated for the following designs:

(1) randomized block,

(2) randomized block design with replication within units,

(3) cross-over,

(4) latin square,

(5) split-plot.

Two observations may have certain subscripts in common. This property enables one to describe the relation between them. Observations related in a specific way will be called associates.

The error sum of squares of the p missing values is a quadratic function, say, $F(x_1, x_2, \ldots, x_p)$ which is given by

$$2(x'A - q') / N.$$

The values of x_i are obtained by equating each of p partial derivatives to zero. These are, in matrix notation

$$Ax = q.$$

In a randomized block design, the error term in the analysis of variance is given by

$$\Sigma_{(i)}\Sigma_{(j)}\, x_{ij}^2 - \Sigma_{(i)}(B_i + \Sigma_{(j)}\, x_{ij})^2 / t - \Sigma_{(j)}(T_j + \Sigma_{(i)}x_{ij})^2/r$$

$$+ (G + \Sigma_{(i)}\Sigma_{(j)}x_{ij})^2/rt + \text{constant.}$$

Taking the partial differential of the above equation with respect to a particular missing observation, x_{gh}, and equating the derivative to zero, gives the equation

$$(r-1)(t-1)x_{gh} + (1-r) \overset{\text{i associates}}{\sum_{\substack{(j) \\ j \neq h}}} x_{gj} + (1-t) \overset{\text{j associates}}{\sum_{\substack{(i) \\ i \neq g}}} x_{ih} + \overset{\text{zero associates}}{\sum_{\substack{(i) \\ i \neq g}} \sum_{\substack{(j) \\ j \neq h}}} x_{ij} = rB_g + tT_h - G \, .$$

This equation can be represented in matrix form as

$$Ax = q \, .$$

The elements of matrix A are given by:

(i) $(r-1)(t-1)$ for the missing observations under consideration,

(ii) $(1-r)$ for the block associates,

(iii) $(1-t)$ for the treatment associates,

(iv) 1 for the zero associates.

One can use the same method, which is described above to get the matrix form $Ax = q$ in other experimental designs.