ANALYSIS OF VARIANCE OF A RANDOMIZED BLOCK DESIGN WITH MISSING OBSERVATIONS

by

William Alexander Glenn

Thesis submitted to the Graduate Faculty of the

Virginia Polytechnic Institute

in candidacy for the degree of

MASTER OF SCIENCE

in

STATISTICS

APPROVED:

Director of Graduate Studies

Dean of Applied Science and Business Administration

APPROVED:

Head of Department

Major Professor

September, 1957

Blacksburg, Virginia

TABLE OF CONTENTS

I. INTRODUCTION

In many avenues of research one encounters the problem of analyzing results which are classified according to two different criteria. These criteria are usually referred to as blocks and treatments. A block is a group of fairly homogeneous experimental units, while a treatment refers to any combination of factors imposed on a single experimental unit. When the number of experimental units in each block is equal to the number of treatments being investigated, and the treatments are assigned at random to the experimental units in each block independently, the experimental layout is called a randomized block design. In this type of design the total variation among the observed values may be subdivided into three parts: (i) variation among blocks, (ii) variation among treatments, and (iii) experimental error. This property, which is equivalent to the complete separability of the classes of factor effects, is known as orthogonality. Blocks and treatments are said to be orthogonal.

In experimental work it sometimes happens that the results of one or more observations are lost. Animals may become sick or die, field plots may be ravaged by some pest, or labels may be mislaid. On occasions there may be sound external evidence for the rejection of a value as unreliable due to a cause not affecting the other values. For example, the yield of a plot which has been trampled by animals or partially washed out may not be considered comparable with that of the other plots. Despite extreme care in conducting the experiment, when the data are summarized there may be blank cells indicating missing observations.

When the original design is accidentally upset, the methods of analysis which are ordinarily applicable must be modified. If there are missing values, the orthogonality property of the design disappears. The procedure of the analysis of variance applicable to orthogonal experiments cannot be used directly on non-orthogonal data. The additivity of the sums of squares computed in the usual fashion does not hold.

The methods of analysis which have been proposed for dealing with the case in which a few values are missing from an otherwise orthogonal experiment depend on the estimation of the yields of missing plots. A method of doing this on a strictly statistical basis was first developed by Allan and Wishart (1930). They derived formulas for the estimation of a single missing yield in a randomized block and in a Latin square experiment. Their methods were extended by Yates (1933) to cover several missing plots in a given experiment. He used the technique of minimizing the error variance obtained when unknowns are substituted for the missing yields, the extension to the case of several missing plots being based on an iterative procedure. He also showed that in a complete analysis of variance using estimated values for the missing plots, the treatment sum of squares is over-estimated but may be corrected by subtracting the bias, for which he gave a generalized formula.

This thesis will deal with the estimation of several missing values in the randomized block design, by minimizing the error sum of squares. Explicit equations for each value missing will be derived for certain cases. A procedure will be given for the completely general case which may prove to be less tedious in application than the iterative method of Yates. A direct method of analysis not requiring a correction for bias in the treatment sum of squares will be demonstrated. Formulas will be given for the bias introduced if an analysis of variance is carried out on the augmented data. These, though equivalent to that of Yates, will be found to differ from the latter slightly in form.

It should be pointed out that the estimation of missing yields does not recover lost information. The loss of several plots from a randomized block experiment results in a proportionate loss of information planned in the design. The only objective of the techniques described herein is the salvaging of as much as possible from that which remains.

II. THEORETICAL BACKGROUND

2.1 Randomized Block Design.

Consider data presented in the following array, in which it is assumed that each y_{ij} is obtained independently of all other y's:

Table 2.1

Notation for Randomized Block Design

	Blocks 1	2	.	j	.	q	Totals
Treatments 1	y_{11}	y_{12}	.	y_{1j}	.	y_{1q}	$Y_{1\cdot}$
2	y_{21}	y_{22}	.	y_{2j}	.	y_{2q}	$Y_{2\cdot}$
.	.	.	.	.	.	.	.
i	y_{i1}	y_{i2}	.	y_{ij}	.	y_{iq}	$Y_{i\cdot}$
.	.	.	.	.	.	.	.
p	y_{p1}	y_{p2}	.	y_{pj}	.	y_{pq}	$Y_{p\cdot}$
Totals	$Y_{\cdot 1}$	$Y_{\cdot 2}$	.	$Y_{\cdot j}$	.	$Y_{\cdot q}$	$Y_{\cdot\cdot}$

where y_{ij} = observation in the i^{th} treatment and j^{th} block,

$$Y_{i\cdot} = \sum_{j=1}^{q} y_{ij},\ Y_{\cdot j} = \sum_{i=1}^{p} y_{ij},$$

and

$$Y_{\cdot\cdot} = \sum_{i=1}^{p} Y_{i\cdot} = \sum_{j=1}^{q} Y_{\cdot j}.$$

The usual analysis of variance assumptions for this type of design may be summarized in the following way: It is assumed that an observation may be represented by the mathematical model

$$y_{ij} = \mu + \gamma_i + \beta_j + \epsilon_{ij}; \quad i = 1,\ldots,p, \qquad (2.1)$$
$$j = 1,\ldots,q,$$

where μ, γ_i and β_j are constants such that

$$\sum_{i=1}^{p} \gamma_i = \sum_{j=1}^{q} \beta_j = 0.$$

μ represents the overall mean effect, while γ_i represents the true effect of the i^{th} treatment measured as a deviation from the mean, and β_j represents the true effect of the j^{th} block measured as a deviation from the mean. The ϵ_{ij}, representing random error effect, are independently and normally distributed with mean 0 and variance σ^2.

If m, t_i and b_j represent the respective estimates of μ, γ_i and β_j, the application of standard least squares procedures leads to the following:

$$m = \frac{Y_{..}}{pq};$$

$$t_i = \frac{Y_{i.}}{q} - m, \quad i = 1,\ldots,p;$$

$$b_j = \frac{Y_{.j}}{p} - m, \quad j = 1,\ldots,q.$$

It may also be shown that the sums of squares associated with the sources of variation in this type of design are given by

T = total sum of squares

$$= \sum_{i=1}^{p}\sum_{j=1}^{q} y_{ij}^2 - \frac{Y_{..}^2}{pq}, \qquad (2.2)$$

B = block sum of squares

$$= \sum_{j=1}^{q} \frac{Y_{\cdot j}^2}{p} - \frac{Y_{\cdot\cdot}^2}{pq} , \qquad (2.3)$$

A = treatment sum of squares

$$= \sum_{i=1}^{p} \frac{Y_{i\cdot}^2}{q} - \frac{Y_{\cdot\cdot}^2}{pq} , \qquad (2.4)$$

E = experimental error sum of squares

$$= \sum_{i=1}^{p} \sum_{j=1}^{q} (y_{ij} - \frac{Y_{i\cdot}}{q} - \frac{Y_{\cdot j}}{p} + \frac{Y_{\cdot\cdot}}{pq})^2$$

$$= T - B - A. \qquad (2.5)$$

In order to test the hypothesis that there is no significant effect due to treatments ($H_0 : \tau_i = 0, \ i = 1,\ldots,p$) a table is usually set up in the following form:

Table 2.2

Analysis of Variance

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F
Blocks	q-1	B	B/(q-1)	
Treatments	p-1	A	A/(p-1)	*
Error	(p-1)(q-1)	E	E/(p-1)(q-1)	
Total	pq - 1	T		

The test of H_0 is effected by comparing the ratio of mean square for treatments to mean square for experimental error (entered in the table as shown: *) with the tabular value of the F distribution with $(p-1)$ and $(p-1)(q-1)$ degrees of freedom at a chosen significance level, and drawing the appropriate conclusion.

2.2 Missing Observations.

Consider data as presented in Table 2.1 with the exception that a few (one or more) of the y_{ij} values are missing. The cases in which a complete block or a complete treatment is missing are not being considered here since they present no computational difficulties. When a few observations are missing two approaches are available to the research worker: (a) the standard least squares procedure applied to the known data or (b) the estimation of values to replace the missing observations followed by the analysis of Section 2.1 applied to the augmented data, with adjustments to the degrees of freedom and treatment sum of squares which will be described below. While (a) is the correct procedure, the least squares normal equations which result will lack symmetry due to the absence of terms corresponding to missing values. Attention is thus turned to (b) as an approach which is simpler to apply. It is desired, of course, that the application of (b) reproduce the results which would be given by (a).

In (a) values of m, t_i and b_j are so chosen as to minimize

$$E' = \sum_{\substack{\text{known} \\ \text{data}}}\sum (y_{ij} - m - t_i - b_j)^2 \qquad (2.6)$$

where the prime on E' distinguishes this error term from E in Section 2.1 where no data are missing. Suppose now that unknowns x_{ij} are inserted for the missing values. A yield value will thus be represented by y_{ij} if present, and by x_{ij} if absent. The criterion to be used in determining the values of the x_{ij} is that of minimizing the experimental error sum of squares taken over all data, observed and estimated. This technique was proposed by Yates (1933), following a suggestion from R.A. Fisher, and is currently favored by statisticians. Let E^* represent the experimental error sum of squares when the data are augmented by the insertion of the values of the x_{ij}. Let m^*, t_i^* and b_j^* be the least squares estimates which minimize E^*. Then

$$E^* = \sum_{\substack{\text{known}\\ \text{data}}}\sum (y_{ij} - m^* - t_i^* - b_j^*)^2 + \sum_{\substack{\text{missing}\\ \text{plots}}}\sum (x_{ij} - m^* - t_i^* - b_j^*)^2, \quad (2.7)$$

Differentiating (2.7) with respect to any x_{ij} gives

$$\frac{\partial E^*}{\partial x_{ij}} = 2(x_{ij} - m^* - t_i^* - b_j^*),$$

which is then equated to zero to yield

$$x_{ij} = m^* + t_i^* + b_j^*, \quad (2.8)$$

clearly corresponding to a minimum of E^*. When the value of x_{ij} from (2.8) is inserted in (2.7) the latter becomes

$$E^* = \sum_{\substack{\text{known}\\ \text{data}}}\sum (y_{ij} - m^* - t_i^* - b_j^*)^2. \quad (2.9)$$

Comparison of (2.9) and (2.6) shows that m^*, t_i^* and b_j^* which minimize E^* and m, t_i and b_j which minimize E' are such that

$$m^* = m;\ t_i^* = t_i;\ b_j^* = b_j. \quad (2.10)$$

It is concluded from (2.10), (2.9) and (2.6) that

$$E^* = E' \quad (2.11)$$

These results were first established by Yates. They may be stated as follows: If values of the unknown observations satisfying (2.8) are inserted, and the augmented data analyzed in the manner of Section 2.1, then the following properties hold: (i) the estimates of treatment and block effects are the same as those obtained by the correct least squares procedure and (ii) the error sum of squares is the same as given by the correct procedure.

If n observations are absent the number of independent values in the data is n less than if there were none missing. Thus the total number of degrees of freedom is reduced by n. Unless one or more complete blocks or treatments is missing, the number of parameters required to describe these effects will not be changed. It follows that the missing degrees of freedom all come from the error sum of squares. Therefore in making a test of significance using augmented data the degrees of freedom for total and error must each be reduced by the number of values inserted.

The formula for a single missing value given by Yates follows from (2.8). Suppose that the observation for treatment h, block k is missing. Let

$Y'_{h\cdot}$ = total of observed values in treatment h,

$Y'_{\cdot k}$ = total of observed values in block k

and $Y'_{\cdot\cdot}$ = total of all observed values.

Then representing the missing observation by x_{hk} and applying the theory of Section 2.1 to the augmented data one obtains

$$m^* = \frac{Y'_{\cdot\cdot} + x_{hk}}{pq},\ t^*_h = \frac{Y'_{h\cdot} + x_{hk}}{q} - m^*,\ b^*_k = \frac{Y'_{\cdot k} + x_{hk}}{p} - m^*.$$

Insertion of these values in (2.8) yields

$$x_{hk} = \frac{Y'_{h\cdot} + x_{hk}}{q} + \frac{Y'_{\cdot k} + x_{hk}}{p} - \frac{Y'_{\cdot\cdot} + x_{hk}}{pq},$$

which is then solved for x_{hk} to give

$$x_{hk} = \frac{pY'_{h\cdot} + qY'_{\cdot k} - Y'_{\cdot\cdot}}{(p-1)(q-1)}, \qquad (2.12)$$

in agreement with Yates' result.

2.3 Bias in the Treatment Sum of Squares.

In one important respect the method of inserting a value or values and analyzing the completed data gives a result which fails to agree with that of the correct least squares procedure. The treatment sum of squares obtained in the analysis of the augmented data is always larger than the correct treatment sum of squares. This bias tends to enhance the apparent significance in the F test.

The existence of the bias may be established by showing that the variance of the mean of any treatment in which an augmenting value satisfying (2.8) has been inserted exceeds the variance expected if all data are present. This expected value is σ^2/q since there are q columns, σ^2 having been defined in Section 2.1. Suppose that observation (h,k) is missing and the value given by (2.12) is inserted for it. In order to find $V(Y_{h\cdot}/q)$, the variance of the mean of treatment h, first write

$$Y_{h\cdot} = Y'_{h\cdot} + x_{hk}.$$

By substitution from (2.12) this may be written as

$$Y_{h.} = Y'_{h.} + \frac{pY'_{h.} + qY'_{.k} - Y'_{..}}{(p-1)(q-1)} = \frac{(pq-q)Y'_{h.} + (q-1)Y'_{.k} - (Y'_{..} - Y'_{h.} - Y'_{.k})}{(p-1)(q-1)} ,$$

from which it is seen that the variance of $Y_{h.}$ is given by

$$V(Y_{h.}) = \frac{q^2(p-1)^2(q-1)\sigma^2 + (q-1)^2(p-1)\sigma^2 + (p-1)(q-1)\sigma^2}{\left[(p-1)(q-1)\right]^2} .$$

After some algebraic manipulation one obtains

$$V(Y_{h.}) = \frac{q^2\sigma^2}{(p-1)(q-1)}\left[(p-1) + \frac{1}{q}\right] = q\,\sigma^2\left[1 + \frac{p}{(p-1)(q-1)}\right] .$$

Thus the variance of the mean of treatment h is given by

$$V\left[\frac{Y_{h.}}{q}\right] = \frac{V(Y_{h.})}{q^2} = \frac{\sigma^2}{q}\left[1 + \frac{p}{(p-1)(q-1)}\right] .$$

Since $p/(p-1)(q-1)$ must be positive, the variance of the augmented treatment mean exceeds its expectation of σ^2/q, or the treatment mean square when an augmenting value has been inserted is greater than the correct treatment mean square.

Yates has stated the procedure for finding the correct treatment sum of squares in a randomized block experiment with missing plots. One takes the difference of the sum of squares removed by the fitting of constants corresponding to both blocks and treatments and that removed by constants for blocks only. The sum of squares removed by block and treatment constants is found by calculating the total sum of squares of the original yields without missing values (less correction for the mean) and deducting the residual sum of squares found by analysis of the augmented data. The latter was shown (2.11) to be equivalent to the residual sum of squares obtained in the correct least squares procedure. The sum of squares removed by blocks only is obtained directly from the original block totals by following the procedure of the analysis of variance for the one-way classification with unequal numbers in the classes (blocks).

Let T^*, B^*, A^* and E^* refer to sums of squares (corrected for the mean) associated with total, blocks, treatments and error respectively when the procedure of Section 2.1 is carried out on the augmented data. Let T', B', A' and E' refer to the corresponding quantities when the correct least squares procedure is carried out on the known data. Let $Y'_{h\cdot}$, $Y'_{\cdot k}$, $Y'_{\cdot\cdot}$ represent totals of actually observed values, as in Section 2.2. The procedure outlined above may be expressed in this notation. One calculates from the augmented data

$$E^* = T^* - B^* - A^*$$

$$= E' \text{ by } (2.11)$$

and T' from the original data, to obtain by subtraction

$$B' + A' = T' - E' = T' - T^* + B^* + A^* .$$

Calculating B' from the original block totals and subtracting it from the above, one obtains

$$A' = T' - T^* + B^* + A^* - B' \qquad (2.13)$$

as an expression for the correct treatment sum of squares. A formula for the bias in A^* may be obtained from (2.13) as

$$\text{Bias} = A^* - A' = T^* - T' + B' - B^*. \qquad (2.14)$$

In using (2.14) to express the bias in terms of quantities in the data table, it is convenient to distinguish between the following two cases: (a) when no two of the missing values are in the same block and (b) when more than one value is missing in a given block.

Case (a).

Suppose that n values are missing, no two being in the same block. Let x_{ij} represent estimates satisfying (2.8) where i,j take on the values corresponding to the missing plots. Then one has

$$T^* = \sum_{\substack{\text{known}\\\text{data}}}\sum y_{ij}^2 + \sum_{\substack{\text{missing}\\\text{plots}}}\sum x_{ij}^2 - \frac{\left[Y'_{..} + \sum\limits_{\substack{\text{missing}\\\text{plots}}}\sum x_{ij}\right]^2}{pq},$$

$$T' = \sum_{\substack{\text{known}\\\text{data}}}\sum y_{ij}^2 - \frac{Y'^2_{..}}{pq-n},$$

$$B' = \sum_{\substack{\text{known}\\\text{data}}} \frac{Y^2_{\cdot j}}{p} + \sum_{\substack{\text{missing}\\\text{plots}}} \frac{Y'^2_{\cdot j}}{p-1} - \frac{Y'^2_{..}}{pq-n},$$

$$B^* = \sum_{\substack{\text{known}\\\text{data}}} \frac{Y^2_{\cdot j}}{p} + \sum_{\substack{\text{missing}\\\text{plots}}}\sum \frac{(Y'_{\cdot j} + x_{ij})^2}{p} - \frac{\left[Y'_{..} + \sum\limits_{\substack{\text{missing}\\\text{plots}}}\sum x_{ij}\right]^2}{pq}.$$

Combining these in accord with (2.14) one obtains

$$\text{Bias} = \sum_{\substack{\text{missing}\\\text{plots}}}\sum \left[x_{ij}^2 + \frac{Y'^2_{\cdot j}}{p-1} - \frac{(Y'_{\cdot j} + x_{ij})^2}{p}\right].$$

After some algebraic manipulation this is expressed as

$$\text{Bias} = \frac{1}{p(p-1)} \sum_{\substack{\text{missing}\\\text{plots}}}\sum \left[Y'_{\cdot j} - (p-1)x_{ij}\right]^2. \quad (2.15)$$

Case (b).

Suppose that n values are missing, all in the k^{th} block. Let x_{ik} represent estimates satisfying (2.8) where i takes on values corresponding to the missing plots. Then one has

$$T^* = \sum_{\substack{\text{known}\\ \text{data}}}\sum y_{ij}^2 + \sum_{\substack{\text{missing}\\ \text{plots}}} x_{ik}^2 - \frac{\left[Y'_{..} + \sum_{\substack{\text{missing}\\ \text{plots}}} x_{ik}\right]^2}{pq},$$

$$T' = \sum_{\substack{\text{known}\\ \text{data}}}\sum y_{ij}^2 - \frac{Y'^2_{..}}{pq-n},$$

$$B' = \sum_{j\neq k} \frac{Y^2_{.j}}{p} + \frac{Y'^2_{.k}}{p-n} - \frac{Y'^2_{..}}{pq-n},$$

$$B^* = \sum_{j\neq k} \frac{Y^2_{.j}}{p} + \frac{\left[Y'_{.k} + \sum_{\substack{\text{missing}\\ \text{plots}}} x_{ik}\right]^2}{p} - \frac{\left[Y'_{..} + \sum_{\substack{\text{missing}\\ \text{plots}}} x_{ik}\right]^2}{pq}.$$

Combining these in accord with (2.14) one obtains

$$\text{Bias} = \sum_{\substack{\text{missing}\\ \text{plots}}} x_{ik}^2 + \frac{Y'^2_{.k}}{p-n} - \frac{\left[Y'_{.k} + \sum_{\substack{\text{missing}\\ \text{plots}}} x_{ik}\right]^2}{p}. \qquad (2.16)$$

Clearly this formula holds only if $n < p$, which is reasonable since if n were equal to p the entire k^{th} block would be missing.

Formulas (2.15) and (2.16) in combination are easily shown to be equivalent to the generalized formula given by Yates. The terms which they involve are readily available in the data table when an analysis is being performed. For a single missing value either one of them reduces to the formula covering that particular case.

To determine the bias in a general case one should proceed as follows: Using (2.15) sum over all blocks with one missing value; apply (2.16) with $n = 2$ to all blocks having two missing values; apply (2.16) with $n = 3$ to all blocks having three missing values, and so on. The total bias is the sum of the terms thus obtained.

2.4 A Direct Method of Analysis.

The direct calculation of the bias may be avoided by using a method based on the theory of Section 2.3. The estimates for the missing yields (determination of which is covered later in this thesis) are first entered in the table. The analysis of the augmented data is then carried out as in Section 2.1. However, in recording the sums of squares for total and blocks (before correction for the mean) a separation into two parts is made for later reference. In the case of total sum of squares one part corresponds to values originally present, the other part to values inserted. In the case of blocks one part gives the sum of squares for complete blocks, while the other part arises from blocks in which values were inserted. No separation is necessary in the treatment sum of squares. A test of significance of treatment effect may be made on the augmented data in the regular manner, degrees of freedom being adjusted as in Section 2.2. Should this approximate test prove non-significant, one may stop there without fear of having drawn a wrong conclusion. This is because the bias always enhances the apparent significance, as shown in Section 2.3.

If, however, the approximate test indicates significance, one proceeds to find the correct treatment sum of squares to be tested against error. This is accomplished by deleting the inserted values and treating the data as a one-way classification with unequal numbers in the blocks. The work is facilitated by referring to the separation in sums of squares mentioned above. The total sum of squares (before correction for the mean) has been calculated as the part corresponding to values originally present. The block sum of squares is found by adding to the value already obtained for complete blocks the part corresponding to blocks having missing plots. Note must be taken of the actual number in the block in performing the divisions. The appropriate correction factor for the mean based on data originally present is then subtracted from each of the above. The correct treatment sum of squares is found by subtracting from the total sum of squares for the one-way classification the block sum of squares for the one-way classification and the error sum of squares obtained from the augmented data. For this calculation Yates has used a table of the form shown below. The notation is that of Section 2.3.

Table 2.3

Calculation of Correct Treatment Sum of Squares

Total (based on known data):	T'
Error (from augmented data):	$E^* = E'$
Difference = Blocks + Treatments:	$T' - E' = B' + A'$
Blocks (based on known data):	B'
Difference = Correct Treatment Sum of Squares:	A'

III. ESTIMATION OF SEVERAL MISSING VALUES

3.0 Introduction.

For the estimation of several missing values in a randomized block experiment (or other standard design) Yates (1933) has proposed an iterative procedure. First of all, approximate values (such as the mean of all recorded observations) are inserted for all except one of the missing yields. This one is then estimated with the formula (2.12) for a single missing observation. One at a time in succession the approximate values are removed, treatment, block and grand totals adjusted, and (2.12) used to estimate a value for each in turn until a complete set is obtained. The entire process is repeated, using as the approximate values the estimates obtained on the previous iteration. This is continued until the values obtained on a given iteration are respectively the same as those of the stage before. When such is the case the desired values have been found. Yates has pointed out that this method is equivalent to the solution by iterations of the system of linear equations obtained by partially differentiating the error sum of squares with respect to the unknowns representing missing yields, and setting the derivatives equal to zero.

The criterion of minimizing the error sum of squares when an analysis of variance is performed on the augmented data has resulted in equation (2.8). The condition is seen to be satisfied if one inserts for each missing value the quantity corresponding to its expectation under least squares analysis. Otherwise stated, the value inserted is such as to make a zero contribution to the error sum of squares, accounting for the property expressed in (2.11).

If n values are missing and the unknown for each one is set equal to its own expected value, a system of simultaneous linear equations results. In certain cases the system leads to a general solution for all permissible values of n. These cases are when no two of the values missing are in the same block or treatment and when all of the missing values are in a single block or treatment, to be covered in the sections below. A discussion of the general case will be given in IV.

3.1 Case of n Missing Values, no Two in the Same Block or Treatment.

Let the missing values be represented by x_{cd}, x_{ef}, x_{gu}, ..., n in all, where $c \neq e \neq g \neq \ldots$ and $d \neq f \neq u \neq \ldots$. Let no treatment subscript exceed p and no block subscript exceed q. Replacing each x by its expected value given by (2.8), applying the theory of Section 2.1 to the augmented data and using notation introduced in Section 2.2 one obtains

$$x_{cd} = \frac{Y'_{c\cdot} + x_{cd}}{q} + \frac{Y'_{\cdot d} + x_{cd}}{p} - \frac{Y'_{\cdot\cdot} + x_{cd} + x_{ef} + x_{gu} + \cdots\cdots}{pq},$$

$$x_{ef} = \frac{Y'_{e\cdot} + x_{ef}}{q} + \frac{Y'_{\cdot f} + x_{ef}}{p} - \frac{Y'_{\cdot\cdot} + x_{cd} + x_{ef} + x_{gu} + \cdots\cdots}{pq},$$

$$x_{gu} = \frac{Y'_{g\cdot} + x_{gu}}{q} + \frac{Y'_{\cdot u} + x_{gu}}{p} - \frac{Y'_{\cdot\cdot} + x_{cd} + x_{ef} + x_{gu} + \cdots\cdots}{pq},$$

$$\cdot\cdot \qquad \cdot\cdot \qquad \cdot\cdot \qquad \cdot\cdot \qquad .$$

After collecting the x's and multiplying through by pq in each equation, one expresses the system in the form

$$
\begin{array}{cccccc}
x_{cd}(p-1)(q-1) + & x_{ef} + & x_{gu} & + \cdots = & pY'_{c\cdot} + qY'_{\cdot d} - Y'_{\cdot\cdot}\,, \\
x_{cd} & + x_{ef}(p-1)(q-1) + & x_{gu} & + \cdots = & pY'_{e\cdot} + qY'_{\cdot f} - Y'_{\cdot\cdot}\,, \\
x_{cd} & + x_{ef} & + x_{gu}(p-1)(q-1) & + \cdots = & pY'_{g\cdot} + qY'_{\cdot u} - Y'_{\cdot\cdot}\,, \\
\cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot
\end{array}
$$

Defining the quantity z as follows:

$$z_{ij} = pY'_{i\cdot} + qY'_{\cdot j} - Y'_{\cdot\cdot}\,, \qquad (3.1)$$

one may express the above system in matrix notation as

$$\begin{bmatrix} (p-1)(q-1) & 1 & \cdot & 1 \\ 1 & (p-1)(q-1) & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 1 & \cdot & \cdot & (p-1)(q-1) \end{bmatrix} \begin{bmatrix} x_{cd} \\ x_{ef} \\ x_{gu} \\ \cdots \end{bmatrix} = \begin{bmatrix} z_{cd} \\ z_{ef} \\ z_{gu} \\ \cdots \end{bmatrix} . \qquad (3.2)$$

The symmetric n x n matrix appearing in (3.2) is seen to have $(p-1)(q-1)$ as each diagonal element and ones elsewhere. Its inverse is readily found to be the symmetric n x n matrix such that each diagonal element is

$$\frac{(p-1)(q-1) + (n-2)}{(p-1)^2(q-1)^2 + (n-2)(p-1)(q-1) - (n-1)}\,,$$

and every other element is

$$\frac{-1}{(p-1)^2(q-1)^2 + (n-2)(p-1)(q-1) - (n-1)}\,.$$

The solution of (3.2) is thus given by

$$x_{hk} = \frac{\left[(p-1)(q-1) + (n-2)\right] z_{hk} - \underset{\substack{\text{missing plots} \\ i\neq h,\ j\neq k}}{\sum\sum} z_{ij}}{(p-1)^2(q-1)^2 + (n-2)(p-1)(q-1) - (n-1)}\,, \qquad (3.3)$$

where h,k take on successively the pairs of values corresponding to missing plots. The summation in the numerator refers to the sum of z values for all missing plots except hk.

3.2 Case of n Missing Values, All in a Single Block or Treatment.

Consider first the case in which n values are missing in treatment r where $n < q$, so that there is at least one value present in the treatment. Let the missing values be represented by x_{rc}, x_{rd}, x_{re}, ... where no block subscript exceeds q and r does not exceed p. As in Section 3.1 a system of linear equations is obtained. This system is expressed in matrix notation as

$$\begin{bmatrix} (p-1)(q-1) & (1-p) & \cdot & (1-p) \\ (1-p) & (p-1)(q-1) & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ (1-p) & \cdot & \cdot & (p-1)(q-1) \end{bmatrix} \begin{bmatrix} x_{rc} \\ x_{rd} \\ x_{re} \\ \cdots \end{bmatrix} = \begin{bmatrix} z_{rc} \\ z_{rd} \\ z_{re} \\ \cdots \end{bmatrix} . \quad (3.4)$$

The symmetric n x n matrix appearing in (3.4) has $(p-1)(q-1)$ as each diagonal element and $(1-p)$ elsewhere. Its inverse is readily found to be the symmetric n x n matrix such that each diagonal element is

$$\frac{q - n + 1}{(p-1)\left[(q-1)^2 - (n-2)(q-1) - (n-1)\right]} ,$$

and every other element is

$$\frac{1}{(p-1)\left[(q-1)^2 - (n-2)(q-1) - (n-1)\right]} .$$

The formula for the estimation of <u>n missing values, all in treatment r</u> <u>$(n < q)$</u> is therefore

$$x_{rk} = \frac{(q - n + 1)z_{rk} + \sum\limits_{\substack{\text{missing plots} \\ j \neq k}} z_{rj}}{(p-1)\left[(q-1)^2 - (n-2)(q-1) - (n-1)\right]}, \qquad (3.5)$$

where k takes on successively the values corresponding to missing plots in treatment r. The summation in the numerator refers to the sum of z values for all missing plots in treatment r other than rk.

In the same manner one obtains the formula for the estimation of n missing values, all in block s $(n < p)$ as

$$x_{hs} = \frac{(p - n + 1)z_{hs} + \sum\limits_{\substack{\text{missing plots} \\ i \neq h}} z_{is}}{(q-1)\left[(p-1)^2 - (n-2)(p-1) - (n-1)\right]}, \qquad (3.6)$$

where h takes on successively the values corresponding to missing plots in block s. The summation in the numerator refers to the sum of z values for all missing plots in block s other than hs. In deriving (3.6) it is found that the matrix in the equation corresponding to (3.4) differs from the matrix in the latter only in the fact that p and q are interchanged.

In these two cases the restrictions $n < q$ and $n < p$ are introduced since for $n = q$ (3.5) breaks down and similarly (3.6) for $n = p$. This means that there must be at least one value present in the treatment or block to which one proposes to apply these formulas. It is clear, however, that if only one treatment appeared in a block the block should be dropped, since the one plot can convey no information on treatment differences.

3.3 Estimation of Two Missing Values.

If only two values are missing, they must necessarily fall into one of the three categories discussed in Sections 3.1 and 3.2. The appropriate formulas are obtained from (3.3), (3.5) and (3.6) respectively.

If two values are missing, not in the same block or treatment, let them be represented by x_{cd} and x_{ef} where $c \neq e$, $d \neq f$, the larger of c and e does not exceed p, and the larger of d and f does not exceed q. Then (3.3) gives

$$x_{cd} = \frac{(p-1)(q-1)z_{cd} - z_{ef}}{(p-1)^2(q-1)^2 - 1}, \qquad x_{ef} = \frac{(p-1)(q-1)z_{ef} - z_{cd}}{(p-1)^2(q-1)^2 - 1}.$$

If two values are missing, both in treatment r, let them be represented by x_{rc} and x_{rd}, where the larger of c and d does not exceed q and r does not exceed p. Then (3.5) gives

$$x_{rc} = \frac{(q-1)z_{rc} + z_{rd}}{(p-1)\left[(q-1)^2-1\right]}, \qquad x_{rd} = \frac{(q-1)z_{rd} + z_{rc}}{(p-1)\left[(q-1)^2-1\right]}.$$

If two values are missing, both in block s, let them be represented by x_{cs} and x_{ds}, where the larger of c and d does not exceed p and s does not exceed q. Then (3.6) gives

$$x_{cs} = \frac{(p-1)z_{cs} + z_{ds}}{(q-1)\left[(p-1)^2-1\right]}, \qquad x_{ds} = \frac{(p-1)z_{ds} + z_{cs}}{(q-1)\left[(p-1)^2-1\right]}.$$

3.4 Estimation of Three Missing Values.

When three values are missing in a randomized block design, they must occur in one of the following arrangements:

Case 1: no two in the same block or treatment,

Case 2: all three in the same block or treatment,

Case 3: value rs missing, together with another value from treatment r, and another from block s,

Case 4: two values missing in treatment r, together with a third not in treatment r, nor in the same block as either of the two,

Case 5: two values missing in block s, together with a third not in block s, nor in the same treatment as either of the two.

Cases 1 and 2 are covered by formulas (3.3), (3.5) and (3.6) with $n = 3$. Formulas are derived below for each of the other cases.

For Case 3 let the missing values be represented by x_{rc}, x_{rs} and x_{ds} where $d \neq r$, $c \neq s$, no treatment subscript exceeds p and no block subscript exceeds q. Replacing each x by its expectation given by (2.8), applying the theory of Section 2.1 to the augmented data and using notation introduced in Section 2.2, one obtains

$$x_{rc} = \frac{Y'_{r\cdot} + x_{rc} + x_{rs}}{q} + \frac{Y'_{\cdot c} + x_{rc}}{p} - \frac{Y'_{\cdot\cdot} + x_{rc} + x_{rs} + x_{ds}}{pq} ,$$

$$x_{rs} = \frac{Y'_{r\cdot} + x_{rc} + x_{rs}}{q} + \frac{Y'_{\cdot s} + x_{rs} + x_{ds}}{p} - \frac{Y'_{\cdot\cdot} + x_{rc} + x_{rs} + x_{ds}}{pq} ,$$

$$x_{ds} = \frac{Y'_{d\cdot} + x_{ds}}{q} + \frac{Y'_{\cdot s} + x_{rs} + x_{ds}}{p} - \frac{Y'_{\cdot\cdot} + x_{rc} + x_{rs} + x_{ds}}{pq} .$$

After collecting the x's and multiplying through by pq in each equation, one obtains a system of equations which may be expressed in matrix notation as

$$\begin{bmatrix} (p-1)(q-1) & (1-p) & 1 \\ (1-p) & (p-1)(q-1) & (1-q) \\ 1 & (1-q) & (p-1)(q-1) \end{bmatrix} \begin{bmatrix} x_{rc} \\ x_{rs} \\ x_{ds} \end{bmatrix} = \begin{bmatrix} z_{rc} \\ z_{rs} \\ z_{ds} \end{bmatrix}, \quad (3.7)$$

where z_{ij} has been defined in (3.1). It will be convenient to introduce the following notation:

$$v = 1-p, \quad w = 1-q, \quad (3.8)$$

in terms of which (3.7) assumes the form

$$\begin{bmatrix} vw & v & 1 \\ v & vw & w \\ 1 & w & vw \end{bmatrix} \begin{bmatrix} x_{rc} \\ x_{rs} \\ x_{ds} \end{bmatrix} = \begin{bmatrix} z_{rc} \\ z_{rs} \\ z_{ds} \end{bmatrix}. \quad (3.9)$$

The inverse of the symmetric 3 x 3 matrix in (3.9) is found to be

$$\frac{1}{vw(v^2-1)(w^2-1)} \begin{bmatrix} w^2(v^2-1) & -w(v^2-1) & 0 \\ -w(v^2-1) & v^2w^2-1 & -v(w^2-1) \\ 0 & -v(w^2-1) & v^2(w^2-1) \end{bmatrix}.$$

The solution of (3.9) is thus given by

$$x_{rc} = \frac{wz_{rc} - z_{rs}}{v(w^2-1)}, \quad (3.10)$$

$$x_{rs} = \frac{(v^2w^2-1)z_{rs}}{vw(v^2-1)(w^2-1)} - \frac{z_{rc}}{v(w^2-1)} - \frac{z_{ds}}{w(v^2-1)} \quad (3.11)$$

and

$$x_{ds} = \frac{vz_{ds} - z_{rs}}{w(v^2-1)}. \quad (3.12)$$

In (3.10), (3.11) and (3.12) one has the solution for Case 3 expressed in terms of notation introduced in (3.1) and (3.8). When applying these formulas one should note that (3.11) refers to the value x_{rs}

while (3.10) refers to the missing value in the same treatment as x_{rs} and (3.12) to the missing value in the same block as x_{rs}.

For Case 4 let the missing values be x_{rc}, x_{rd} and x_{ef} where $e \neq r$, $c \neq d \neq f$, no treatment subscript exceeds p and no block subscript exceeds q. Proceeding in a manner similar to that for Case 3 one obtains the following system of equations:

$$x_{rc} = \frac{Y'_{r\cdot} + x_{rc} + x_{rd}}{q} + \frac{Y'_{\cdot c} + x_{rc}}{p} - \frac{Y'_{\cdot\cdot} + x_{rc} + x_{rd} + x_{ef}}{pq},$$

$$x_{rd} = \frac{Y'_{r\cdot} + x_{rc} + x_{rd}}{q} + \frac{Y'_{\cdot d} + x_{rd}}{p} - \frac{Y'_{\cdot\cdot} + x_{rc} + x_{rd} + x_{ef}}{pq},$$

$$x_{ef} = \frac{Y'_{e\cdot} + x_{ef}}{q} + \frac{Y'_{\cdot f} + x_{ef}}{p} - \frac{Y'_{\cdot\cdot} + x_{rc} + x_{rd} + x_{ef}}{pq}.$$

Performing the same simplifying operations as in Case 3 one expresses the above system in matrix form as

$$\begin{bmatrix} (p-1)(q-1) & (1-p) & 1 \\ (1-p) & (p-1)(q-1) & 1 \\ 1 & 1 & (p-1)(q-1) \end{bmatrix} \begin{bmatrix} x_{rc} \\ x_{rd} \\ x_{ef} \end{bmatrix} = \begin{bmatrix} z_{rc} \\ z_{rd} \\ z_{ef} \end{bmatrix}. \quad (3.13)$$

In terms of the notation introduced in (3.8) the above assumes the form

$$\begin{bmatrix} vw & v & 1 \\ v & vw & 1 \\ 1 & 1 & vw \end{bmatrix} \begin{bmatrix} x_{rc} \\ x_{rd} \\ x_{ef} \end{bmatrix} = \begin{bmatrix} z_{rc} \\ z_{rd} \\ z_{ef} \end{bmatrix}. \quad (3.14)$$

The inverse of the symmetric 3 x 3 matrix in (3.14) is found to be

$$\frac{1}{v^2w^2 + v^2w-2}\begin{bmatrix} \frac{v^2w^2-1}{v(w-1)} & \frac{1-v^2w}{v(w-1)} & -1 \\ \frac{1-v^2w}{v(w-1)} & \frac{v^2w^2-1}{v(w-1)} & -1 \\ -1 & -1 & v(w+1) \end{bmatrix}.$$

The solution of (3.14) is thus given by

$$x_{rc} = \frac{1}{v^2w^2 + v^2w-2}\left[\frac{v^2w^2-1}{v(w-1)} z_{rc} + \frac{1-v^2w}{v(w-1)} z_{rd} - z_{ef}\right], \quad (3.15)$$

$$x_{rd} = \frac{1}{v^2w^2 + v^2w-2}\left[\frac{v^2w^2-1}{v(w-1)} z_{rd} + \frac{1-v^2w}{v(w-1)} z_{rc} - z_{ef}\right] \quad (3.16)$$

and $$x_{ef} = \frac{1}{v^2w^2 + v^2w-2}\left[v(w+1)z_{ef} - (z_{rc} + z_{rd})\right]. \quad (3.17)$$

In (3.15), (3.16) and (3.17) one has the solution for Case 4 expressed in terms of notation introduced in (3.1) and (3.8). When applying these formulas it should be noted that (3.15) and (3.16) refer to the missing values in treatment r, while (3.17) refers to the value not in the same block or treatment as either of the first two.

In case 5 let the missing values be represented by x_{cs}, x_{ds} and x_{ef} where $f \neq s$, $c \neq d \neq e$, no treatment subscript exceeds p and no block subscript exceeds q. Proceeding as in Case 3, one obtains a system of equations which, when simplified as in previous cases assume the form

$$\begin{bmatrix} (p-1)(q-1) & (1-q) & 1 \\ (1-q) & (p-1)(q-1) & 1 \\ 1 & 1 & (p-1)(q-1) \end{bmatrix}\begin{bmatrix} x_{cs} \\ x_{ds} \\ x_{ef} \end{bmatrix} = \begin{bmatrix} z_{cs} \\ z_{ds} \\ z_{ef} \end{bmatrix}. \quad (3.18)$$

In terms of the notation introduced in (3.8) the above assumes the form

$$\begin{bmatrix} vw & w & 1 \\ w & vw & 1 \\ 1 & 1 & vw \end{bmatrix} \begin{bmatrix} x_{cs} \\ x_{ds} \\ x_{ef} \end{bmatrix} = \begin{bmatrix} z_{cs} \\ z_{ds} \\ z_{ef} \end{bmatrix} . \qquad (3.19)$$

The coefficient matrix in (3.18) differs from that in (3.13) only in the fact that the roles of p and q are interchanged. Similarly the matrix in (3.19) is obtained from that in (3.14) by interchanging the roles of v and w. It follows that the solution of (3.19) is given by

$$x_{cs} = \frac{1}{w^2v^2 + w^2v - 2}\left[\frac{w^2v^2 - 1}{w(v-1)} z_{cs} + \frac{1 - w^2v}{w(v-1)} z_{ds} - z_{ef}\right], \qquad (3.20)$$

$$x_{ds} = \frac{1}{w^2v^2 + w^2v - 2}\left[\frac{w^2v^2 - 1}{w(v-1)} z_{ds} + \frac{1 - w^2v}{w(v-1)} z_{cs} - z_{ef}\right] \qquad (3.21)$$

and $$x_{ef} = \frac{1}{w^2v^2 + w^2v - 2}\left[w(v+1)z_{ef} - (z_{cs} + z_{ds})\right]. \qquad (3.22)$$

The solution for Case 5 is thus expressed in terms of notation introduced in (3.1) and (3.8). When applying these formulas it should be noted that (3.20) and (3.21) refer to the missing values in block s, while (3.22) refers to the value not in the same block or treatment as either of the first two.

IV. GENERAL CASE

When n missing values in a randomized block design are distributed arbitrarily over the blocks and treatments, it is not possible to give an explicit equation for each value missing, covering all permissible values of n. The derivations in III and the results obtained for the cases considered there make this fact evident. However, it is possible to set down certain features which apply in general.

When the criterion of minimizing the error sum of squares is applied, the unknowns (x's) representing missing values are obtained as the solutions of a system of linear equations which may be expressed in matrix notation as

$$U\,X = Z \qquad (4.1)$$

where U is an n x n matrix to be described below, X is an n x 1 column vector of unknowns representing missing values and Z is an n x 1 column vector of corresponding z values as defined by (3.1). The cases discussed in III have illustrated this. In the cases considered there the solution was effected by inverting the U matrix to give a result of the form

$$X = U^{-1}\,Z. \qquad (4.2)$$

The particular structure of the U matrix depends on the manner in which the missing values are distributed relative to one-another over the blocks and treatments. To facilitate discussion of this, consider the case of n missing values and let the vacant plots be numbered in any order from 1 to n. Let the unknown representing the i^{th} missing value in this numbering system be x_i. This x_i plays the same part as the previous x_{ij}, of course, but the single-subscript designation is more convenient for present purposes. The n x 1 column vector X is thus made up of elements x_i where i takes on integral values from 1 to n. This enables the giving of a complete description of the n x n matrix U, which is always symmetric. The nature of the equations from which it results indicates that (i) every diagonal element is $(p-1)(q-1)$; (ii) if x_i and x_j are not in the same block or treatment, element ij is 1; (iii) if x_i and x_j are in the same treatment, element ij is $(1-p)$; (iv) if x_i and x_j are in the same block, element ij is $(1-q)$. Having written the column vector X, one may construct the matrix U from the above properties by referring to the data table. The column vector Z, consisting of elements determined from (3.1) for each vacant plot, must be written in the order which corresponds to that of the x_i's. The system (4.1) is thus obtained.

In order to put the system (4.1) in the solved form (4.2) it is necessary to invert the matrix U. The procedure proposed by Yates (1933) has been described as being equivalent to solving the system by iterations. However, since the matrix U is symmetric, its inversion by a method such as the abbreviated Doolittle may well prove to be less troublesome than the iterative method. The two methods will give precisely the same results, on condition that the iterative procedure is carried through a sufficient number of stages.

V. EXAMPLES

Illustrations, given below, of the application of the techniques described above are based on a randomized block experiment conducted by G. D. Jones at Orange, Virginia in 1955. The experiment was concerned with the investigation of phosphorus fertilization for alfalfa. The yields are presented in Table 5.1.

Table 5.1

Total Annual Yield of Alfalfa in Pounds of Air-dried Hay

	Blocks 1	2	3	4	5	6	Totals
Treatments 1	13.75	18.30	22.39	23.72	21.86	23.60	123.62
2	13.05	22.69	21.73	24.29	23.27	23.23	128.26
3	18.67	17.51	18.38	22.83	23.84	23.61	124.84
4	15.36	17.45	18.39	20.97	21.31	23.86	117.34
5	19.13	22.16	23.42	26.12	20.75	28.52	140.10
6	21.87	21.34	23.73	23.53	27.84	21.93	140.24
Totals	101.83	119.45	128.04	141.46	138.87	144.75	774.40

The six treatments were described as follows:

1. 500 pounds of P_2O_5 on the plow sole,
2. 500 pounds of P_2O_5 plowed under,
3. 500 pounds of P_2O_5 top dressed,
4. 100 pounds of P_2O_5 top dressed annually,

5. 500 pounds of P_2O_5 on the plow sole together with 100 pounds of P_2O_5 top dressed annually,

6. 1000 pounds of P_2O_5 on the plow sole.

All plots received 50 pounds of P_2O_5 at seeding in 1952, and 100 pounds of K_2O annually. The plot size was 150 square feet.

The results of analyzing the data of Table 5.1 in the manner of Section 2.1 are presented in Table 5.2. The calculated value of $F = 3.03$ is compared with the tabular value of the F distribution with (5,25) degrees of freedom at, say, the .05 level. Since the latter is 2.60, one rejects the hypothesis that there is no significant effect due to treatments, the Type I error being 5%.

Table 5.2

Analysis of Variance for Data in Table 5.1

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F
Blocks	5	221.8396	44.3679	
Treatments	5	72.0457	14.4091	3.03
Error	25	119.0381	4.7615	
Total	35	412.9234		

Consider now the appropriate procedure when certain of the yield values of Table 5.1 are absent. For example, suppose that the observations for treatment 5, block 1, treatment 5, block 4 and treatment 6, block 4 are missing. This is immediately identified as an example of Case 3 considered in Section 3.4. Accordingly let the missing values be represented by x_{51}, x_{54} and x_{64}. It is evident from Table 5.1 that, in the notation of Section 2.2

$$Y'_{5\cdot} = 94.85,\ Y'_{6\cdot} = 116.71,\ Y'_{\cdot 1} = 82.70,\ Y'_{\cdot 4} = 91.81,\ Y'_{\cdot\cdot} = 705.62.$$

Using the above with $p = q = 6$ in (3.1) yields

$$z_{51} = 359.68,\ z_{54} = 414.34 \text{ and } z_{64} = 545.50.$$

According to (3.8)

$$v = w = -5.$$

Thus one obtains from (3.10)

$$x_{51} = \frac{-5(359.68) - 414.34}{-5(24)} = 18.44,$$

from (3.11)

$$x_{54} = \frac{(624)(414.34)}{25(24)(24)} + \frac{359.68}{120} + \frac{545.50}{120} = 25.50$$

and from (3.12)

$$x_{64} = \frac{-5(545.50) - 414.34}{-5(24)} = 26.18.$$

These values are inserted in their positions in the table and the analysis of the augmented data is carried out in the manner of Section 2.1. In doing the calculations the separation of sums of squares explained in Section 2.4 is observed. The results are

$$\frac{Y^2_{\cdot\cdot}}{pq} = \frac{(775.74)^2}{36} = 16{,}715.9041,$$

$$T^* = 15{,}469.2556 + 1675.6760 - 16{,}715.9041 = 429.0275,$$

$$B^* = \frac{70{,}899.9835}{6} + \frac{30{,}818.6797}{6} - 16{,}715.9041 = 237.2064,$$

$$A^* = \frac{100{,}766.4494}{6} - 16{,}715.9041 = 78.5041$$

$$E^* = T^* - B^* - A^* = 113.3170.$$

The approximate test described in Section 2.4 is set up as shown in Table 5.3. Since three values have been inserted the degrees of freedom for total and error have each been reduced by three in accord with the theory of Section 2.2.

Table 5.3

Approximate Analysis of Variance

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F
Blocks	5	237.2064	47.4413	
Treatments	5	78.5041	15.7008	3.05
Error	22	113.3170	5.1508	
Total	32	429.0275		

Since the approximate test indicates significance at the 5% level, the appropriate tabular value of F being 2.66, one proceeds to find the correct treatment sum of squares to be tested against error. Deleting the inserted values and treating the data as a one-way classification with unequal numbers in the blocks one calculates

$$\frac{Y_{..}'^2}{pq-3} = \frac{(705.62)^2}{33} = 15{,}087.8662,$$

$$T' = 15{,}469.2556 - 15{,}087.8662 = 381.3894,$$

$$B' = \frac{70{,}899.9835}{6} + \frac{6839.2900}{5} + \frac{8429.0761}{4} - 15{,}087.8662 = 203.9247,$$

$$A' = T' - B' - E^{*} = 381.3894 - 203.9247 - 113.3170 = 64.1477.$$

In the above it may be noted that in getting T' use was made of 15,469.2556 previously calculated in connection with T^{*}. Similarly 70,899.9835, the sum of squares related to complete blocks used in getting B^{*} appeared also in B'. The analysis is completed by setting up the exact analysis of variance appearing in Table 5.4.

Table 5.4

Exact Analysis of Variance

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F
Blocks	5	203.9247	40.7849	
Treatments	5	64.1477	12.8295	2.49
Error	22	113.3170	5.1508	
Total	32	381.3894		

When the above calculated F is compared with 2.66, the appropriate tabular value at the 5% level, it is concluded that the effect due to treatments is not significant. This means that the loss of the particular three values in this example has changed the conclusion in the F test from significance to non-significance. It is further noted that the bias in this case is sufficient to change the conclusion from significance in the approximate test to non-significance in the exact test.

The above method of analysis avoids the direct use of a formula for the bias. However, should one wish to calculate the bias separately, it may be obtained by the use of formulas (2.15) and (2.16). Use of (2.15) for block 1 yields

$$\text{Bias} = \frac{[82.70 - 5(18.44)]^2}{6(5)} = 3.0083,$$

while use of (2.16) with n = 2 for block 4 yields

$$\text{Bias} = (25.50)^2 + (26.18)^2 + \frac{(91.81)^2}{4} - \frac{(143.49)^2}{6} = 11.3480.$$

Taking the sum of these gives the total bias as 14.3563. It is noted that this agrees with the difference between the augmented and correct treatment sums of squares, apart from a rounding error of one unit in the fourth decimal place.

It is of interest to consider the extent to which the F ratio may be affected when estimates are inserted for missing values. In order to illustrate this, certain values in Table 5.1 have been deleted and estimates inserted by the procedures described above. Pertinent details of the results are presented in Table 5.5. The first column indicates the values which were considered missing and estimated. The second column gives the resulting bias in the treatment sum of squares, while the third column gives the correct treatment sum of squares. The fourth and fifth columns give the calculated F ratios on the basis of the approximate and exact tests respectively. In each case three values were considered missing. The appropriate comparison values from the F table for (5,22) degrees of freedom are 2.66 at the 5% level and

3.99 at the 1% level.

Table 5.5

Results of the Estimation of Certain Missing Values

Values Estimated	Bias	Correct Treatment Sum of Squares	F Ratio on Approximate Test	F Ratio on Exact Test
x_{14}, x_{24}, x_{35}	2.5054	76.4835	3.07	2.97
x_{54}, x_{56}, x_{65}	5.9974	36.5828	1.93	1.66
x_{41}, x_{56}, x_{65}	8.7784	39.4865	2.15	1.76
x_{46}, x_{56}, x_{66}	15.1166	77.6263	4.64	3.88
x_{62}, x_{64}, x_{66}	57.5040	104.5170	9.10	5.87

The first set in the above table is an example of Case 5, considered in Section 3.4. The values deleted were from treatments whose means did not differ significantly. The bias is small and the test of significance is not seriously affected. The second set is an example of Case 4 of Section 3.4, while the third is of the type considered in Section 3.1. In both of these cases the values deleted came from the treatments which had the most marked effect in producing the significant difference originally. Loss of these values is seen to have had the effect of considerably reducing the treatment sum of squares, while the error mean square was found to have dropped only slightly as compared to the value which it had when all data were present, resulting in a reduced F ratio. The fourth and fifth sets are examples of the type considered in Section 3.2. In the fourth a relatively small error mean square accounted for the higher F ratio. It is noted in this case that a conclusion of significance at the 1% level could be drawn if one were to rely solely on the approximate test, but such is not the case for the exact test. In the fifth case three values were deleted from the treatment with the highest mean. The result was a markedly reduced error mean square and an enhanced treatment mean square, accounting for the large calculated F ratios. The bias in this case amounts to 35% of the augmented treatment sum of squares.

VI. SUMMARY

The estimation of several missing values in a randomized block design is considered. The method used is that of minimizing the error sum of squares, proposed originally by Yates (1933). Explicit equations for each absent value are derived for all cases in which not more than three values are missing. A general formula valid for any permissible number of missing observations is given for the case in which no two values are missing in the same block or treatment, and also for the case in which all of the values missing are in a single block or treatment. A procedure for the completely general case is proposed. This, although requiring the inversion of a symmetric matrix of order equal to the number of missing observations, may prove to be less tedious in application than the iterative method proposed by Yates.

A direct method of analysis not requiring a correction for bias in the treatment sum of squares is discussed and demonstrated. Formulas are given for the bias introduced when an analysis of variance is carried out on the augmented data. These, though equivalent to the generalized formula given by Yates, are found to differ from the latter slightly in form.

Examples of the application of the techniques described are given. One of these is shown in detail, while for others, the general procedure being the same, only the pertinent conclusions are presented.

VII. ACKNOWLEDGMENTS

The author wishes to express his sincere appreciation to Dr. Clyde Y. Kramer for his guidance during the course of this work. His many suggestions and willing assistance have made this paper possible. An expression of gratitude is extended also to Dr. Boyd Harshbarger for his attentive interest, encouragement and aid in the field of statistics.

VIII. BIBLIOGRAPHY

1. Allan, F. E., and Wishart, J. (1930). "A Method of Estimating the Yield of a Missing Plot in Field Experimental Work." Journal of Agricultural Science, Vol. 20, pp. 399-406.

2. Cochran, W. G., and Cox, G. M. (1950). Experimental Designs. John Wiley and Sons, Inc., New York.

3. Ostle, B. (1954). Statistics in Research. The Iowa State College Press, Ames, Iowa.

4. Yates, F. (1933). "The Analysis of Replicated Experiments When the Field Results are Incomplete." Empire Journal of Experimental Agriculture, Vol. 1, pp. 129-142.

The vita has been removed from the scanned document