AN INVESTIGATION INTO MISSING OBSERVATIONS IN A RANDOMIZED BLOCK EXPERIMENT

A THESIS

PRESENTED TO

THE FACULTY OF GRADUATE STUDIES

UNIVERSITY OF MANITOBA

IN PARTIAL FULFILLMENT

OF THE REQUIREMENTS OF THE DEGREE

MASTER OF SCIENCE

BY

MARY LYNN FOLLIOTT

APRIL 1970

ABSTRACT

In experimental research, it is not always possible to obtain the desired information from each experimental unit. This can be due to such things as death, disease or human errors. This paper investigates the estimation and analysis of these experiments. The investigation is done in reference to the randomized block design.

The procedure of estimating the missing data is divided into two cases namely, one missing observation and more than one missing value. For the case of one missing value, either a least squares analysis or a covariance analysis is used to calculate an estimate. When there are more than one missing observation, a procedure presented in 1959 by Biggers is used to obtain the estimates. Using an estimate in an analysis of variance, introduces a bias in the treatment sum of squares. The formulation of this bias is derived for the most general case.

Experimental results generated by a computer are used to obtain the emperical powers of the F values associated with the data with estimates for missing observations. These experimental results are compared with the theoretical values of the power.

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION

A statistical analysis of experimental data obtained from a population with an underlying linear model is carried out under the following assumptions, 1) randomness of the errors, 2) their mutual independence, 3) normality and 4) homogeneity of variance. In most of these situations, it is necessary to have the observed values for each experimental unit. In those cases when one or more observations are missing, a value must be supplied to these experimental units so that the standard analysis can be performed. The methods of estimating values for the missing data, the effect of these estimates on the analysis of the experiment, and the power of the statistical test used in the analysis will be investigated in this paper. This will be done in reference to the randomized block model.

In the case of one missing observation, least squares analysis and covariance analysis are used to calculate the required estimate. Both methods are theoretically identical, but differ in ease of application. For manual calculations, the least squares analysis is used but when

the calculations are done by a computer, a covariance analysis is performed. The covariance method lends itself more easily to programming. When more than one observation is missing, a procedure developed by Biggers (1959) is used to obtain the essential estimates.

After supplying the required estimates, an analysis of variance performed on the data will result in an upward bias in the treatment sum of squares. The formula given by Kenney and Keeping (1951) is valid only for the case when there is at most one missing value in a treatment or block group. The general formula for which this bias is a special case will be derived in chapter III.

Emperical results, generated by an I.B.M. 360/65, are used to obtain the experimental or Monte Carlo power of the F-test associated with the treatment sum of squares. The power is found by counting the proportion of F values greater than the critical value, in a set of K independent identically distributed F values. This estimate of the power can be compared with the theoretical power obtained from tables given by Tiku (1967).

Allan and Wishart (1930) were the first to develop a formula which estimated a value for a missing value. In 1933, Yates demonstrated that Allan and Wishart had obtained their formula by minimizing the sum of squares due to error with respect to the missing observation. Yates also presented an iterative procedure for calculating

estimates when there are several missing values.

Tables for the power of the F-test were published by Tang (1938). He developed a set of tables which depended on the degrees of freedom of the F-test and a function ϕ of the non-centrality parameter λ. Pearson and Hartley (1951) and Fox (1956) charted the function ϕ dependent on the degrees of freedom and for several values of α the probability of the Type I error and β the probability of the Type II error. Tiku (1967) extended Tang's tables and also gave a method of interpolation to obtain a value of the power not given in the tables.

CHAPTER II

ONE MISSING VALUE

It sometimes happens that one or more observations in an experiment are missing. In such cases, an estimate for each missing value must be obtained. If an analysis is performed with an experimental unit missing, the sum of squares for treatments, blocks and error do not add up to the total sum of squares. In other words, the model is not additive. This can be shown in an example as follows. Given a randomized block design with 3 treatments and 3 blocks, and a missing observation in treatment 2 block 3,

BLOCK	TREATMENT			TOTAL	MEAN
	1	2	3		
1	9	3	9	21	7
2	8	5	2	15	5
3	4		10	14	7
TOTAL	21	8	21	50	$6\frac{1}{4}$
MEAN	7	4	7	$6\frac{1}{4}$	

The analysis is made omitting observation x_{32}.

$$TSS = 9^2 + 3^2 + \ldots + 4^2 + 10^2 - \frac{(50)^2}{8} = 67\tfrac{1}{2}$$

$$SST = \frac{(21)^2}{3} + \frac{(8)^2}{2} + \frac{(21)^2}{3} - \frac{(50)^2}{8} = 7\tfrac{1}{2}$$

$$SSB = \frac{(21)^2}{3} + \frac{(15)^2}{3} + \frac{(14)^2}{2} - \frac{(50)^2}{8} = 13\tfrac{1}{2}$$

$$SSE = TSS - SST - SSB = 47\tfrac{1}{2}$$

The calculation of the error sum of squares is given by.

$$SSE = (9-7-7+6\tfrac{1}{4})^2 + (3-7-4+6\tfrac{1}{4})^2 + \ldots + (10-7-7+6\tfrac{1}{4})^2 = 49\tfrac{1}{2}$$

SSE (subtraction) $\neq$ SSE (calculated).

Two familiar methods of calculating a missing value are,

1) least squares analysis,

2) covariance analysis.

In least squares analysis, we begin with the mathematical model of the randomized block design.

$$x_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \qquad \begin{matrix} i=1,2,\ldots,t \\ j=1,2,\ldots,b \end{matrix}$$

where x_{ij} is defined in terms of a general mean μ, a treatment effect α_i, a block effect β_j and a random error ε_{ij}. The values of μ, α_i and β_j are determined so as to minimize the sum of squares due to error. The sum of squares due to error if found to be equal to,

$$\Sigma_i\Sigma_j x_{ij}^2 - \frac{\Sigma_i(\Sigma_j x_{ij})^2}{b} - \frac{\Sigma_j(\Sigma_i x_{ij})^2}{t} + \frac{(\Sigma_i\Sigma_j x_{ij})^2}{tb}$$

When a value is missing an estimate can be found by minimizing the sum of squares due to error with respect to this value. By obtaining an estimate in this way, a value is found which contributes least to the error sum of squares.

To obtain the value which is to be assigned to the missing experimental unit consider the following in which x_{gh} is missing.

$$SSE = \Sigma_i\Sigma_j x_{ij}^2 - \frac{\Sigma_i(\Sigma_j x_{ij})^2}{b} - \frac{\Sigma_j(\Sigma_i x_{ij})^2}{t} + \frac{(\Sigma_i\Sigma_j x_{ij})^2}{tb}$$

$$\frac{\partial}{\partial x_{gh}} SSE = x_{gh} - \frac{\Sigma_j x_{gj}}{b} - \frac{\Sigma_i x_{ih}}{t} + \frac{\Sigma_i\Sigma_j x_{ij}}{tb}$$

Set this equal to zero and solve for x_{gh}.

$$x_{gh}\left(1 - \frac{1}{b} - \frac{1}{t} + \frac{1}{tb}\right) + \left(-\frac{\Sigma_{j\neq h} x_{gj}}{b} - \frac{\Sigma_{i\neq g} x_{ih}}{t} + \frac{\Sigma_{i\neq g}\Sigma_{j\neq h} x_{ij}}{tb}\right) = 0$$

$$x_{gh}\left(\frac{(b-1)(t-1)}{tb}\right) = \frac{\left(tT'_g + bB'_h - G'\right)}{tb}$$

$$x_{gh} = \frac{\left(tT'_g + bB'_h - G'\right)}{(t-1)(b-1)}$$

where

$$T'_g = \Sigma_j x_{gj} - x_{gh}$$

$$B'_h = \Sigma_i x_{ih} - x_{gh}$$

$$G' = \Sigma_i\Sigma_j x_{ij} - x_{gh}$$

Having an estimate of x_{gh}, the analysis of variance can be carried out in the usual manner. However, the treatment sum of squares when calculated using this estimate is biased. By this we mean that the expected value of the treatment mean squares is equal to

$$\sigma^2 + \frac{b}{(t-1)} \Sigma_i \alpha_i^2 + q.$$

Under the null hypothesis H_o: $\alpha_i=0$, this value becomes

$$\sigma^2 + q', \text{ where } q'>0.$$

The usual F-test gives rise to an exaggeration of the treatment effect.

The amount of bias can be found by considering the hypothesis that there are no differences between treatments ie. H_o: $\alpha_i=0$ for all i. By combining the treatment and error sums of squares and minimizing this conditional error denoted by S_c with respect to the missing value,a different estimate for the missing observation will be obtained. This is demonstrated by,

$$S_c = \Sigma_i \Sigma_j x_{ij}^{*2} - \frac{\Sigma_j (\Sigma_i x_{ij}^*)^2}{t}$$

$$\frac{\partial}{\partial x_{gh}^*} S_c = x_{gh}^* - \frac{\Sigma_i x_{ih}^*}{t}$$

Set equal to zero and solve for x_{gh}^*.

$$x_{gh}^* (1 - \frac{1}{t}) = \frac{B_h'}{t}$$

$$x^{*}_{gh} = \frac{B'_h}{(t-1)} = A$$

The bias is introduced by using

$$x_{gh} = \frac{tT'_g + bB'_h - G'}{(t-1)(b-1)}$$

as an estimate instead of

$$x_{gh} = \frac{B'_h}{(t-1)}$$

in testing the null hypothesis. The former should be used in order to obtain an unbiased estimate of the missing value. The amount of bias is,

$$x_{gh}^2 - A^2 - \frac{(x_{gh} - B'_h)^2}{t} + \frac{(A - B'_h)^2}{t}$$

$$= \frac{(t-1)}{t}(x_{gh}^2 - A^2) - \frac{2B'_h}{t}(x_{gh} - A)$$

$$= \frac{(t-1)}{t}(x_{gh} - A)(x_{gh} + A - 2A)$$

$$= \frac{(t-1)}{t}(x_{gh} - A)^2$$

$$= \frac{(t-1)}{t}(x_{gh} - \frac{B'_h}{(t-1)})^2$$

Considering the degrees of freedom associated with the error sum of squares, each observation x_{ij} has an associated ε_{ij}. It is assumed that $\Sigma_i \varepsilon_{ij} = \Sigma_j \varepsilon_{ij} = 0$. Besides this, it is known that $\varepsilon_{gh}=0$ since observation x_{gh}

is missing. Evaluating the number of dependent ε_{ij}'s, there are t treatments and b blocks, hence there are t+b-1 dependent ε_{ij}'s. Besides this there is one due to ε_{gh} being equal to zero. As a result, there will be tb-(t+b-1+1) or tb-t-b = (t-1)(b-1)-1 independent ε_{ij}'s and the degrees of freedom associated with the sum of squares due to error is one less than the case with no missing values.

Consider estimating the value of the observation missing in the example at the beginning of this chapter ie. x_{32}.

$$x_{32} = \frac{3(8) + 3(14) - 50}{(2)(2)} = 4$$

ANOVA

SOURCE OF VARIATION	D.F.	SS	MS	F
Treatments(adj.)	2	18-6=12	6	0.375
Blocks	2	6	3	
Error	4-1=3	48	16	
Total	8-1=7	72-6=66		

$$\text{bias} = \frac{2}{3}(4 - \frac{14}{2})^2 = 6$$

In covariance analysis the mathematical model for a randomized block design is given by,

$$y_{ij} = \mu + \alpha_i + \rho_j + \beta x_{ij} + \varepsilon_{ij}$$
$$i=1,2,\ldots,t$$
$$j=i,2,\ldots,p$$

where y_{ij} is the dependent variable given in terms of a general mean μ, a treatment effect α_i, a block effect ρ_j,

a β-multiple of the independent variable x_{ij}, and a random error ε_{ij}. The underlying assumptions of the model are,

1) the x's are fixed variables and measured without error.

2) the ε's are independent normally distributed variables with mean zero and common variance σ^2.

3) the regression of y on x after the removal of the treatment and block effects is linear and independent of treatments and blocks.

When an observation is missing, an estimate of its value is found by performing a covariance with the dependent y-variates being the existing observations with zero for the missing value, and the independent x-variates consisting of zeros corresponding to the existing observations and a minus one for the missing observation. The value of the regression coefficient b is the estimate used for the the missing value. This can be shown by the following, where x_{gh} is assumed to be missing.

$$\hat{y}_{ij} = \bar{y}_{i.} + \bar{y}_{.j} - \bar{\bar{y}}_{..} + b(x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{\bar{x}}_{..})$$

hence

$$y_{gh} = \bar{y}_{g.} + \bar{y}_{.h} - \bar{\bar{y}}_{..} + b(x_{gh} - \bar{x}_{g.} - \bar{x}_{.h} + \bar{\bar{x}}_{..})$$

now

$$\bar{y}_{g.} = \frac{T'_g}{p} \qquad \bar{y}_{.h} = \frac{B'_h}{t} \qquad \bar{\bar{y}}_{..} = \frac{G'}{tp}$$

$$\hat{y}_{gh} = \frac{T'_g}{p} + \frac{B'_h}{t} - \frac{G'}{tp} + b(x_{gh} - \bar{x}_{g.} - \bar{x}_{.h} + \bar{\bar{x}}_{..})$$

also

$$b = \frac{tT'_g + pB'_h - G'}{(t-1)(p-1)}$$

for $x_{gh} = -1$,

$$\hat{y}_{gh} = \frac{tT'_g + pB'_h - G'}{tp} + \left(\frac{tT'_g + pB'_h - G'}{(t-1)(p-1)}\right)\left(-1 + \frac{1}{p} + \frac{1}{t} - \frac{1}{tp}\right)$$

$$= \frac{tT'_g + pB'_h - G'}{tp} + \left(\frac{tT'_g + pB'_h - G'}{(t-1)(p-1)}\right)\left(-\frac{(t-1)(p-1)}{tp}\right)$$

$$= 0$$

for $x_{gh} = 0$,

$$\hat{y}_{gh} = \frac{tT'_g + pB'_h - G'}{tp} + \left(\frac{tT'_g + pB'_h - G'}{(t-1)(p-1)}\right)\left(0 + \frac{1}{p} + \frac{1}{t} - \frac{1}{tp}\right)$$

$$= \frac{tT'_g + pB'_h - G'}{tp} + \left(\frac{tT'_g + pB'_h - G'}{(t-1)(p-1)}\right)\left(-1 + \frac{1}{p} + \frac{1}{t} - \frac{1}{tp} + 1\right)$$

$$= 0 + \frac{tT'_g + pB'_h - G'}{(t-1)(p-1)}$$

$$= b$$

This shows that for x_{gh} equal to minus one, the corresponding y-variate will be zero. If instead of minus one, the value of x_{gh} had been zero (as was the case for the other observations), the corresponding y value would be the regression coefficient b. Hence the estimate used for the missing y_{gh} is b.

A complete analysis of covairiance table for the design is given below.

ANCOVA

SOURCE	D.F.	SUMS OF PROD. XX	XY	YY	D.F.	ADJ. ΣYY
TOTAL	tp-1	Σxx	Σxy	Σyy		
BLOCKS	p-1	B_{xx}	B_{xy}	B_{yy}		
TREAT.	t-1	T_{xx}	T_{xy}	T_{yy}		
ERROR	(t-1)(p-1)	E_{xx}	E_{xy}	E_{yy}	(t-1)(p-1)-1	$E_{yy}-\frac{E_{xy}^2}{E_{xx}}$
T + E	p(t-1)	S_{xx}	S_{xy}	S_{yy}	p(t-1)-1	$S_{yy}-\frac{S_{xy}^2}{S_{xx}}$
T(ADJ)					t-1	$S_{yy}-\frac{S_{xy}^2}{S_{xx}}-E_{yy}+\frac{E_{xy}^2}{E_{xx}}$

Notation:

$$\Sigma xx = \Sigma_i \Sigma_j (x_{ij} - \bar{\bar{x}}_{..})^2 = \frac{tp-1}{tp} \qquad \Sigma yy = \Sigma_i \Sigma_j (y_{ij} - \bar{\bar{y}}_{..})^2$$

$$\Sigma xy = \Sigma_i \Sigma_j (x_{ij} - \bar{\bar{x}}_{..})(y_{ij} - \bar{\bar{y}}_{..}) = \frac{G'}{tp}$$

$$B_{xx} = \frac{\Sigma_j (\bar{x}_{.j} - \bar{\bar{x}}_{..})^2}{t} = \frac{p-1}{tp} \qquad B_{yy} = \frac{\Sigma_j (\bar{y}_{.j} - \bar{\bar{y}}_{..})^2}{t}$$

$$B_{xy} = \frac{\Sigma_j (\bar{x}_{.j} - \bar{\bar{x}}_{..})(\bar{y}_{.j} - \bar{\bar{y}}_{..})}{t} = \frac{G' - pB'_h}{tp}$$

$$T_{xx} = \frac{\Sigma_i (\bar{x}_{i.} - \bar{\bar{x}}_{..})^2}{p} = \frac{t-1}{tp} \qquad T_{yy} = \frac{\Sigma_i (\bar{y}_{i.} - \bar{\bar{y}}_{..})^2}{p}$$

$$T_{xy} = \frac{\Sigma_i (\bar{x}_{i.} - \bar{\bar{x}}_{..})(\bar{y}_{i.} - \bar{\bar{y}}_{..})}{p} = \frac{G' - tT'_g}{tp}$$

$$E_{xx} = \Sigma xx - B_{xx} - T_{xx} = \frac{(t-1)(p-1)}{tp} \qquad E_{yy} = \Sigma yy - B_{yy} - T_{yy}$$

$$E_{xy} = \Sigma xy - B_{xy} - T_{xy} = \frac{tT'_g + pB'_h - G'}{tp}$$

$$S_{xx} = T_{xx} + E_{xx} = \frac{t-1}{t} \qquad S_{yy} = T_{yy} + E_{yy}$$

$$S_{xy} = T_{xy} + E_{xy} = \frac{B'_h}{t}$$

Comparing the two methods of estimating a missing value, it is noticed that both methods obtain the same estimates for the missing observation. Both analyses decrease the degrees of freedom associated with the error sum of squares by one. The adjustment to the treatment sum of squares can be seen by considering the following, Notation:

$$SSE = E_{yy} - \frac{E_{xy}^2}{E_{xx}}$$

$$SS(T+E) = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

$$SST(ADJ) = SS(T+E) - SSE$$

$$T_{yy} = \frac{\Sigma_i y_{i.}^2}{p} - \frac{(\Sigma_i \Sigma_j y_{ij})^2}{tp}$$

T^*_{yy} is the sum of squares for treatments with the missing y observation having its value of zero replaced by the regression coefficient b.

$$T_{yy}^{*} = \frac{(\Sigma_{i\neq g}\bar{y}_{i.}^{2} + (\bar{y}_{g.} + b)^{2})}{p} - \frac{(\Sigma_{i}\Sigma_{j}(y_{ij} + b))^{2}}{tp}$$

$$SS(T+E) = \Sigma yy - B_{yy} - \frac{(\Sigma xy - B_{xy})^{2}}{(\Sigma xx - B_{xx})}$$

$$= \Sigma yy - B_{yy} - \frac{(\frac{G'}{tp} - \frac{G'}{tp} + \frac{pB_{h}'}{tp})^{2}}{\frac{tp-1}{tp} - \frac{t-1}{tp}}$$

$$= \Sigma yy - B_{yy} - \frac{(B_{h}')^{2}}{t(t-1)}$$

$$SSE = \Sigma yy - B_{yy} - T_{yy} - \frac{(\Sigma xy - B_{xy} - T_{xy})^{2}}{\Sigma xx - B_{xx} - T_{xx}}$$

$$= \Sigma yy - B_{yy} - T_{yy} - \frac{(tT_{g}' + pB_{h}' - G')^{2}}{tp(t-1)(p-1)}$$

$$SST(ADJ) = T_{yy} - \frac{(B_{h}')^{2}}{t(t-1)} + \frac{b(tT_{g}' + pB_{h}' - G')}{tp}$$

$$T_{yy}^{*} = \frac{\Sigma_{i\neq g}\bar{y}_{i.}^{2}}{p} + \frac{\bar{y}_{g.}^{2}}{p} + 2\frac{\bar{y}_{g.}b}{p} + \frac{b^{2}}{p} - \frac{\Sigma_{i}\Sigma_{j}y_{ij}^{2}}{tp}$$

$$- 2\frac{b\Sigma_{i}\Sigma_{j}y_{ij}^{2}}{tp} - \frac{b^{2}}{tp}$$

$$= \frac{\Sigma_{i}y_{i.}^{2}}{p} - \frac{\Sigma_{i}\Sigma_{j}y_{ij}^{2}}{tp} + 2\frac{bT_{g}'}{p} - 2\frac{bG'}{tp} + \frac{b^{2}(t-1)}{tp}$$

$$= T_{yy} + 2\frac{bT_{g}'}{p} - 2\frac{bG'}{tp} + \frac{b^{2}(t-1)}{tp}$$

hence,

$$T_{yy} = T_{yy}^{*} - 2\frac{bT'_g}{p} + 2\frac{bG'}{tp} - \frac{b^2(t-1)}{tp}$$

$$SST(ADJ) = T_{yy}^{*} - 2\frac{bT'_g}{p} + 2\frac{bG'}{tp} - \frac{b^2(t-1)}{tp}$$

$$- \frac{(B'_h)^2}{t(t-1)} + \frac{b(tT'_g + pB'_h - G')}{tp}$$

$$= T_{yy}^{*} - 2\frac{bT'_g}{p} + \frac{bT'_g}{p} + 2\frac{bG'}{tp} - \frac{bG'}{tp}$$

$$- \frac{bB'_h}{t} - \frac{b^2(t-1)}{tp} - \frac{(B'_h)^2}{t(t-1)} + 2\frac{bB'_h}{t}$$

$$= T_{yy}^{*} - \frac{b(tT'_g + pB'_h - G')}{tp} - \frac{b^2(t-1)}{tp}$$

$$+ 2\frac{bB'_h}{t} - \frac{(B'_h)^2}{t(t-1)}$$

$$= T_{yy}^{*} - \frac{b^2(t-1)(p-1)}{tp} - \frac{b^2(t-1)}{tp} + 2\frac{bB'_h}{t} - \frac{(B'_h)^2}{t(t-1)}$$

$$= T_{yy}^{*} - \frac{b^2(t-1)}{t} + 2\frac{bB'_h}{t} - \left(\frac{B'_h}{t-1}\right)^2\frac{(t-1)}{t}$$

$$= T_{yy}^{*} - \frac{(t-1)}{t}\left(b - \frac{B'_h}{t-1}\right)^2$$

Hence the bias is equal to

$$\frac{(t-1)}{t}\left(b - \frac{B'_h}{t-1}\right)^2.$$

This is identical to the bias obtained by least squares

analysis. The two methods, therefore yeild the same results.

Consider the example with observation x_{32} missing. Estimate its value using covariance analysis.

ANCOVA

SOURCE	D.F.	SUMS OF PROD XX	XY	YY	b	ADJUSTED YY D.F.	SS	MS	F
TOTAL	8	$\frac{8}{9}$	$\frac{50}{9}$	$\frac{920}{9}$					
BLOCKS	2	$\frac{2}{9}$	$\frac{8}{9}$	$\frac{86}{9}$					
TREAT.	2	$\frac{2}{9}$	$\frac{26}{9}$	$\frac{338}{9}$					
ERROR	4	$\frac{4}{9}$	$\frac{16}{9}$	$\frac{96}{9}$	4	3	48	16	
T + E	6	$\frac{6}{9}$	$\frac{42}{9}$	$\frac{434}{9}$		5	60		
T(ADJ)						2	12	6	$\frac{3}{8}$

CHAPTER III

MORE THAN ONE MISSING VALUE

When the number of missing observations is greater than one, estimates may be obtained by an iterative procedure suggested by Yates that is equivalent to solving a set of simultaneous equations in K unknowns for K missing values. An alternative procedure is to complete a multiple covariance analysis to obtain a set of regression coefficients which are the estimates for the missing observations. Both of the methods become more and more difficult as the number of missing observations increases. Biggers (1959) presented a procedure which simplifies the calculatio of estimates for the cases with more than one missing value.

The procedure developed by Biggers assumes that there are p independent missing values and the error sum of squares is expressed as a quadratic of these p values. The vector of partial derivatives of this function is given by,

$$2\frac{(\underline{x}'A - \underline{q}')}{N}$$

where,

$\underline{x}$ is a column vector of the p missing values,

A is a p×p symmetric matrix determined by the experimental

design and the missing value,

$\underline{q}$ is a column vector calculated from the available data,

N is a constant determined by the experimental design.

Equating this function to zero and solving for $\underline{x}$.

$$A\underline{x} = \underline{q}$$

$$\underline{x} = A^{-1}\underline{q}$$

Consider a randomized block design with t treatments and b blocks. The sum of squares due to error is given by,

$$\Sigma_i \Sigma_j (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{\bar{x}}_{..})^2$$

$$i=1,2,\ldots,t$$

$$j=1,2,\ldots,b.$$

Evaluating the error sum of squares in terms of summations over the set of missing observations, one obtains

$$SSE = \Sigma_{(i)} \Sigma_{(j)} (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{\bar{x}}_{..})^2 + C$$

$$= \Sigma_{(i)} \Sigma_{(j)} x_{ij}^2 - \Sigma_{(i)} \Sigma_{(j)} \bar{x}_{i.}^2 - \Sigma_{(i)} \Sigma_{(j)} \bar{x}_{.j}^2$$

$$+ \Sigma_{(i)} \Sigma_{(j)} \bar{\bar{x}}_{..}^2 + C$$

$$= \Sigma_{(i)} \Sigma_{(j)} x_{ij}^2 - \frac{1}{b} \Sigma_{(i)} (T_i + \Sigma_{(j)} x_{ij})^2$$

$$- \frac{1}{t} \Sigma_{(j)} (B_j + \Sigma_{(i)} x_{ij})^2 + \frac{1}{tb} (G + \Sigma_{(i)} \Sigma_{(j)} x_{ij})^2 + C$$

where,

$$T_i = \Sigma_j x_{ij} \qquad B_j = \Sigma_i x_{ij} \qquad G = \Sigma_i \Sigma_j x_{ij}$$

and C is a constant made up of those terms that do not contain the missing value. Differentiating the sum of squares due to error with respect to the missing value x_{gh} and equating to zero, we obtain the following,

$$\frac{\partial}{\partial x_{gh}} SSE = x_{gh} - \frac{1}{b}(T_g + \Sigma_{(j)} x_{gj}) - \frac{1}{t}(B_h + \Sigma_{(i)} x_{ih}) + \frac{1}{tb}(G + \Sigma_{(i)}\Sigma_{(j)} x_{ij})$$

$$x_{gh} - \frac{1}{b}\Sigma_{(j)} x_{gj} - \frac{1}{t}\Sigma_{(i)} x_{ih} + \frac{1}{tb}\Sigma_{(i)}\Sigma_{(j)} x_{ij} = \frac{T_g}{b} + \frac{B_h}{t} - \frac{G}{tb}$$

Separating into associates, namely those terms with common i and/or j subscripts,

$$tbx_{gh} - tx_{gh} - bx_{gh} + x_{gh} + (1-t)\Sigma_{(j)\, j\neq h} x_{gj} + (1-b)\Sigma_{(i)\, i\neq g} x_{ih} + \Sigma_{(i)\, i\neq g}\Sigma_{(j)\, j\neq h} x_{ij} = tT_g + bB_h - G$$

$$(t-1)(b-1)x_{gh} + (1-t)\Sigma_{(j)\, j\neq h} x_{gj} + (1-b)\Sigma_{(i)\, i\neq g} x_{ih} + \Sigma_{(i)\, i\neq g}\Sigma_{(j)\, j\neq h} x_{ij} = tT_g + bB_h - G$$

where,

(1-t) is the coefficient associated with the i association ie. treatment associates,

(1-b) corresponds to the j association ie. block associates,

1 is associated with the zero associates,

(t-1)(b-1) is the coefficient associated with the missing observation.

The association matrix is composed of the associations between the missing observation's block and treatment numbers. The q-vector is evaluated as the value of

$$tT_g + bB_h - G$$

for the x_{gh} missing value. Consider the example of a 3×3 randomized block design given in chapter II. Assume observations x_{31} and x are missing. The estimates are found in the following manner.

BLOCK	TREATMENT			TOTAL	MEAN
	1	2	3		
1	9	3		12	6
2	8	5	2	15	5
3	4		10	14	7
TOTAL	21	8	12	41	$5\frac{6}{7}$
MEAN	7	4	6	$5\frac{6}{7}$	

The association matrix A is given by,

$$\begin{pmatrix} 4 & 1 \\ 1 & 4 \end{pmatrix}.$$

A is obtained in the following manner. The block and treatment numbers for each of the missing values is compared

with the block and treatment numbers of the complete set of missing observations. In general, assume that x_{ij} and x_{kl} are the missing values. The comparisons between these values will lead to one of the following four values,

if i = k and j = l then the value is (l-t)(l-b),

if i = k and j $\neq$ l then the value is (l-t),

if i $\neq$ k and j = l then the value is (l-b),

if i $\neq$ k and j $\neq$ l then the value is l.

This value is placed in the association matrix opposite x_{ij} and x_{kl}. In the example under consideration,

x_{31} and x_{31} give the value 4 to the A-matrix,

x_{31} and x_{23} give the value l to the A-matrix,

and x_{23} and x_{23} give the value 4 to the A-matrix.

If instead of observations x_{31} and x_{23} being missing, observations x_{21} and x_{22} were the the missing values. The corresponding association matrix would have been,

$$\begin{pmatrix} 4 & -2 \\ -2 & 4 \end{pmatrix}$$

since x_{21} and x_{22}would have given the value (l-t) = l - 3 = -2

Returning to the example under discussion,

	31	23	T_i	B_j	G
31	4	l	12	12	41
23	l	4	8	14	41
			t=3	b=3	-l

$$\underline{x} = A^{-1}\underline{q}$$

$$\begin{pmatrix} x_{31} \\ x_{23} \end{pmatrix} = \begin{pmatrix} \frac{4}{15} & \frac{-1}{15} \\ \frac{-1}{15} & \frac{4}{15} \end{pmatrix} \times \begin{pmatrix} 31 \\ 25 \end{pmatrix} = \begin{pmatrix} 6\frac{3}{5} \\ 4\frac{3}{5} \end{pmatrix}$$

$$\text{bias} = \frac{2}{3}[\ (6\tfrac{3}{5} - 6)^2 + (4\tfrac{3}{5} - 4)^2\ ]$$

The corrected analysis of variance (CANOVA) is given by,

CANOVA

SOURCES	D.F.	SS	MS	F
TREATS.(ADJ)	2	2.88-0.48=2.40	1.20	0.054
BLOCKS	2	12.58	6.29	
ERROR	4-2=2	44.50	22.25	
TOTAL	8-2=6	59.96-0.48=59.48		

Consider the case where there are more than one missing value in a block. The method of calculating the bias derived in chapter II fails. This can be shown by going back to first principles.

The adjusted sum of squares is found by subtracting the sums of squares due to a combined error and the the error with the estimates substituted in for the missing observations. This combined error term is obtained by finding the difference between the total sum of squares and the sum of squares due to blocks under the assumption that the design has unequal treatment effects in each block. The missing treatment effects are those treatments which are applied to the missing observations.

The bias that results from one missing observation per block was shown to be equal to

$$\frac{(t-1)}{t}(E_j - \frac{B'_j}{t-1})^2$$

where the j^{th} block has a missing observation. The bias is subtracted from the treatment sum of squares for each missing observation. The bias which occurs when there are more than one missing observation per block does not reduce to as compact a formula as in the case of one missing observation.

Notation:

$\{Z_{ij}\} = \{X_{ij}\}$ with missing observations

$\{Y_{ij}\} = \{X_{ij}\}$ with estimates for missing data

$\{E_{ij}\}$ = estimate for a missing value

= 0 for an observed value

N is the number of observations

M is the number of missing observations

K is the number of observed observations $\rightarrow K = N - M$

K_j is the number of observed values in the j^{th} block

$\rightarrow \Sigma_j K_j = K$

t is the number of treatments

b is the number of blocks

$$TSS_z = \Sigma_i \Sigma_j z_{ij}^2 - \frac{(\Sigma_i \Sigma_j z_{ij})^2}{K}$$

$$SSB_z = \Sigma_j \frac{(\Sigma_i z_{ij})^2}{K_j} - \frac{(\Sigma_i \Sigma_j z_{ij})^2}{K}$$

$$SSE_z = \Sigma_i \Sigma_j z_{ij}^2 - \Sigma_j \frac{(\Sigma_i z_{ij})^2}{K_j}$$

$$TSS_y = \Sigma_i \Sigma_j y_{ij}^2 - \frac{(\Sigma_i \Sigma_j y_{ij})^2}{N}$$

$$SSB_y = \Sigma_j \frac{(\Sigma_i y_{ij})^2}{t} - \frac{(\Sigma_i \Sigma_j y_{ij})^2}{N}$$

$$SSE_y = \Sigma_i \Sigma_j y_{ij}^2 - \Sigma_j \frac{(\Sigma_i y_{ij})^2}{t} - SST_y$$

$$SST(ADJ)_y = SSE_z - SSE_y$$

$$BIAS = SST_y - SST(ADJ)_y$$

$$= \Sigma_i \Sigma_j y_{ij}^2 - \Sigma_j \frac{(\Sigma_i y_{ij})^2}{t} - \Sigma_i \Sigma_j z_{ij}^2 + \Sigma_j \frac{(\Sigma_i z_{ij})^2}{K_j}$$

$$\Sigma_i \Sigma_j y_{ij}^2 - \Sigma_i \Sigma_j z_{ij}^2 = \Sigma_i \Sigma_j e_{ij}^2$$

$$\Sigma_j \frac{(\Sigma_i y_{ij})^2}{t} - \Sigma_j \frac{(\Sigma_i z_{ij})^2}{K_j}$$

$$= \Sigma_j \left(\frac{(\Sigma_i (z_{ij} + e_{ij}))^2}{t} - \frac{(\Sigma_i z_{ij})^2}{K_j} \right)$$

$$= \Sigma_j \left(\frac{(\Sigma_i z_{ij})^2}{t} + 2\frac{(\Sigma_i z_{ij})(\Sigma_i e_{ij})}{t} + \frac{(\Sigma_i e_{ij})^2}{t} - \frac{(\Sigma_i z_{ij})^2}{K_j} \right)$$

$$= \Sigma_j \left(\frac{(K_j - t)(\Sigma_i z_{ij})^2}{K_j t} + 2\frac{(\Sigma_i z_{ij})(\Sigma_i e_{ij})}{t} + \frac{(\Sigma_i e_{ij})^2}{t} \right)$$

$$= \Sigma_j \left(-\frac{(t-K_j)(1+K_j-1)(t-1-(t-2))}{t}\frac{(\Sigma_i z_{ij})^2}{K_j^2}\right.$$

$$+ 2\frac{(1+K_j-1)(t-1-(t-2))}{t}\left(\frac{\Sigma_i z_{ij}}{K_j}\right)(\Sigma_i e_{ij})$$

$$\left. + \frac{\Sigma_i e_{ij}^2}{t} + \Sigma_i \Sigma_{r\neq i} e_{ij} e_{rj}\right)$$

$$= \Sigma_j \left(-\frac{(t-K_j)(t-1)}{t}\left(\frac{\Sigma_i z_{ij}}{K_j}\right)^2 - \frac{(t-K_j)(K_j-1)}{t}\left(\frac{\Sigma_i z_{ij}}{K_j}\right)^2\right.$$

$$+ \frac{(t-K_j)(t-2)}{t}\left(\frac{\Sigma_i z_{ij}}{K_j}\right)^2 + 2\frac{(t-1)}{t}\left(\frac{\Sigma_i z_{ij}}{K_j}\right)(\Sigma_i e_{ij})$$

$$- 2\frac{(t-2)}{t}\left(\frac{\Sigma_i z_{ij}}{K_j}\right)(\Sigma_i e_{ij}) + 2\frac{(K_j-1)}{t}\left(\frac{\Sigma_i z_{ij}}{K_j}\right)(\Sigma_i e_{ij})$$

$$\left. + \frac{\Sigma_i e_{ij}^2}{t} + \Sigma_i \Sigma_{r\neq i} e_{ij} e_{rj}\right)$$

To simplify this formula, use the following

$$\text{FACTOR} = -\Sigma_j \left(\frac{(t-K_j)(t-K_j-1)}{t}\left(\frac{\Sigma_i z_{ij}}{K_j}\right)^2 - 2\frac{(t-K_j-1)}{t}\left(\frac{\Sigma_i z_{ij}}{K_j}\right)\right.$$

$$\left. \cdot (\Sigma_i e_{ij}) + \Sigma_i \Sigma_{r\neq i} e_{ij} e_{rj}\right)$$

$$\text{BIAS} = \Sigma_i \Sigma_j e_{ij}^2 - \Sigma_j \left(-\frac{(t-K_j)(t-1)}{t}\left(\frac{\Sigma_i z_{ij}}{K_j}\right)^2\right.$$

$$\left. + 2\frac{(t-1)}{t}\left(\frac{\Sigma_i z_{ij}}{K_j}\right)(\Sigma_i e_{ij}) + \frac{\Sigma_i e_{ij}^2}{t}\right) + \text{FACTOR}$$

$$= \Sigma_j \left(\frac{(t-1)}{t}\Sigma_i e_{ij}^2 - 2\frac{(t-1)}{t}\left(\frac{\Sigma_i z_{ij}}{K_j}\right)(\Sigma_i e_{ij})\right.$$

$$+ \frac{(t-K_j)(t-1)}{t}\left(\frac{\Sigma_i z_{ij}}{K_j}\right)^2) + \text{FACTOR}$$

$$= \Sigma_j \frac{(t-1)}{t}(\Sigma_i e_{ij}^2 - 2\left(\frac{\Sigma_i z_{ij}}{K_j}\right)(\Sigma_i e_{ij})$$

$$(t-K_j)\left(\frac{\Sigma_i z_{ij}}{K_j}\right)^2) + \text{FACTOR}$$

In the p^{th} block there are $(t-K_p)$ missing values and hence $(t-K_p)$ estimates. These estimates make up the set $\{E'_{gp}\}_{g=1}^{t-K_p}$. Hence by the deffinition of $\{E_{ij}\}$, K_p of the E_{ip}'s are zero. Define B'_j as $\Sigma_i x_{ij} - \Sigma_{(i)} x_{ij}$, where $\Sigma_{(i)} x_{ij}$ is the sum of the missing values in the j^{th} block. The bias can now be written as,

$$\text{BIAS} = \Sigma_j \frac{(t-1)}{t}(\Sigma_{r=1}^{t-K_j}(e'_{rj} - \frac{B'_j}{t-K_j})^2) + \text{FACTOR}$$

or

$$\text{BIAS} = \Sigma_j \frac{(t-1)}{t}(\Sigma_{r=1}^{t-K_j}(e'_{rj} - \frac{B'_j}{t-K_j})^2) - \Sigma_j(\frac{(t-K_j)(t-K_j-1)}{t}(B'_j)^2$$

$$- 2\frac{(t-K_j-1)}{t}(B'_j)(\Sigma_i e_{ij}) + \Sigma_i \Sigma_{r\neq i} e_{ij} e_{rj})$$

The bias for the cases with only one observation missing per block is a special case of the bias derived above.

$K_j = t - 1$ or $t - K_j = 1$ hence $t - K_j - 1 = 0$

also $\Sigma_i \Sigma_{r \neq i} e_{ij} e_{rj} = 0$, since e_{rj} does not exist for all r. Therefore the trem FACTOR equals zero. The bias becomes,

$$\text{BIAS} = \Sigma_j \frac{(t-1)}{t} (e'_{j1} - \frac{B'_j}{t-1})^2.$$

CHAPTER IV

ANALYSIS OF DATA

In this chapter the estimation of values for missing observations and the subsequent analysis of the experimental results are dicussed. The computer program that was used to analyse the data is given in Appendix A.

A randomized block design with 3 treatments and 5 blocks is used to illustrate the procedure in calculating missing values, in adjusting for bias and in estimating power. The block effects are represented by the set [-3 -2 0 2 3] and the treatment effects are represented by [-1 0 1]. Random errors were chosen from a normal population having mean equal to zero and a variance of two. Using this information, a set of fifteen observations was generated.

The structural part of the model underlying the observations is

$$X_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$
$$i=1,2,\ldots,t$$
$$j=1,2,\ldots,b$$

An analysis of variance is completed on the data and the F value obtained is stored for future use in the calculation

of the power of the F-test.

In the investigation of the problem of missing values, the following procedure was used

1) select one of the fifteen observations at random and delete it from the experiment.
2) estimate a value for the missing observation by a covariance analtsis.
3) obtain the F value associated with the adjusted treatment effect and store this value for future use for the purpose of calculating the power of the F-test.

The two analyses, the complete case and the one with an estimated value, are repeated K times. Experimental powers are obtained from these two sets of F values.

The procedure and analysis for two missing values is completed in a similar manner. The estimates for the missing observations are obtained by using a method developed by Biggers (1959). This procedure is continued for three, four, five and six missing values, respectively.

The limit on the number of missing values allowed in an experiment is dependent on the degrees of freedom associated with the sum of squares due to error. For each missing observation, the degrees of freedom for error loses one degree of freedom. In order to be able to test for treatment differences, we need at least one degree of freedom for error, but in order to apply the power

tables given by Tiku (1967), the error sum of squares must have at least two degrees of freedom. Hence, in the present example, there can be at most six missing observations.

In the calculation of power, a procedure outlined by Tang (1938) and dicussed in Scheffé (1951) is used. Given that

$$\text{and} \quad \alpha = \Pr(\text{reject } H_o/H_o)$$
$$\beta = \Pr(\text{accept } H_o/H_1)$$

then $$\text{power} = 1 - \beta = \Pr(\text{reject } H_o/H_1).$$

The F-test used in testing for differences among treatment effects in an analysis of variance is given by

$$F_{t-1,(t-1)(b-1)} = \frac{\frac{MST}{E(MST)}}{\frac{MSE}{E(MSE)}}$$

where,

$$MST = \frac{1}{t-1}(\Sigma_i\Sigma_j(\bar{x}_{i.} - \bar{\bar{x}}_{..})^2)$$

$$= \frac{1}{t-1}(\Sigma_i\Sigma_j(\alpha_i + \bar{\varepsilon}_{i.} - \bar{\bar{\varepsilon}}_{..})^2)$$

$$MSE = \frac{1}{(t-1)(b-1)}(\Sigma_i\Sigma_j(x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{\bar{x}}_{..})^2)$$

$$= \frac{1}{(t-1)(b-1)}(\Sigma_i\Sigma_j(\varepsilon_{ij} - \bar{\varepsilon}_{i.} - \bar{\varepsilon}_{.j} + \bar{\bar{\varepsilon}}_{..})^2)$$

$$E(MST) = \frac{1}{t-1}(\Sigma_i\Sigma_j\alpha_i^2 + \Sigma_i b\frac{\sigma^2}{b} - tb\frac{\sigma^2}{tb})$$

$$= \frac{b}{t-1}\Sigma_i\alpha^2 + \sigma^2$$