Department of Statistics

UNIVERSITY OF WISCONSIN

Madison, Wisconsin

Technical Report No. 20

July, 1963

ESTIMATING MISSING VALUES IN UNREPLICATED TWO-LEVEL FACTORIAL AND FRACTIONAL FACTORIAL DESIGNS

Norman R. Draper and David M. Stoneman

This research was supported by the United States Navy through the Office of Naval Research, under Contract Nonr-1202(17), Project NR 042 222.

Estimating Missing Values in Unreplicated Two-Level Factorial and Fractional Factorial Designs

Norman R. Draper and David M. Stoneman

University of Wisconsin

0. Introduction

Cochran and Cox (1957) recommend that for a single performance of a complete factorial design, a missing value be estimated by "minimizing the sum of squares for the interactions that are used as error". In this paper, more specific recommendations are made for estimating one or more missing values when the design is a two level factorial or fractional factorial.

1. The purpose of missing value estimation.

The purpose of estimating missing values is to provide numbers which, when used in the place of the missing data, make analysis easy and provide the same estimates (of selected parameters) that would have been obtained, had the incomplete data been analysed directly. When an experiment has been designed, the analysis of incomplete data is often difficult, whereas the "psuedo-complete" data (with estimated missing values inserted) is easily handled either because of the symmetries of the design, or because any complications in the analysis have been previously resolved. The first of these comments applies for example, to the 2^k factorial designs and also to the important 2^{k-p} fractional factorial designs which were described by Box and Hunter (1961 a,b). In many of the situations in which these designs are used, the analysis of variance technique which can be applied to factorial and fractional factorial designs, may not be appropriate. It may be difficult for the experimenter to decide a priori which interactions, if any, be wishes to use as error, especially when a highly fractionated portion of a full factorial has been used and the alias relationships are such that a main effect is involved in many or perhaps all of the linear combinations of effects which can be estimated from the design. If observations from the design were complete, a likely method of analysis in many industrial situations would be to estimate all main effects and interactions and follow this with a half-normal plot of the effects (Daniel, 1959, describes this). Thus if some observations were missing, the advice of Cochran and Cox (1957) to estimate the missing values by minimizing "the sum of squares for the interactions that are used as error" might be difficult to implement because of the experimenter's idea of the appropriate analysis to make and because the alias structure of the design might make prior choice of "the interactions that are used as error" a practical impossibility. It may also lead to insoluble equations. We now suggest a treatment of this situation which has the virtue that it is very simple to employ. This is an important consideration because, often, the estimation of several missing values leads to such computational difficulty that it might have been better not to estimate the values at all but to carry out a (messy) least squares estimation instead - especially if a computer is available. However even if the experimenter does have a digital computer available, his task is not easy. Since in factorial type designs one can estimate only up to as many effects as there are observations, the experimenter must choose which combinations of effects to estimate and which to sacrifice and a wrong choice will lead to biased estimates, just as would be obtained by an erroneous choice of the effects to be used for error.

2. Procedure.

Suppose we perform a two-level factorial or fractional factorial design in n experiments. Then, if the results are complete, we can estimate the "mean" $\bar{y} = (y_1+y_2+\ldots+y_n)/n$, where $y_1, y_2, \ldots, y_n$ are the n observations and $(n-1)$ other "effects". All of the n estimates will be either of pure effects or, in the case of a fractional factorial, linear combinations of effects. Write $\underline{y}' = (y_1, y_2, \ldots, \ldots, y_n)$ for the vector of observations and let $\underline{\ell}_i$, $i=1,2,\ldots,(n-1)$, be the $(n-1)$ vectors for which the $\underline{\ell}_i'\underline{y}$ provide the correct linear combinations of y_i's for estimating the (n-1) effects (or sets of aliased effects) other than the effect involving the mean. For example in a full 2^3 factorial written in standard order we could write $\underline{\ell}_1' = \frac{1}{4}(-1,1,-1,1,-1,1,-1,1)$, when it is clear that $\underline{\ell}_1'\underline{y}$ would give the "1 effect", the estimate of the effect of the first variable under consideration. These $\underline{\ell}_i$ vectors are all orthogonal to each other, and all are orthogonal to a vector $\underline{e}$ where $n\underline{e}' = (1,1,\ldots,1)$, so that $\underline{e}'\underline{y} = \overline{y}$ is the mean of the observations. It follows that the sum of squares due to the n estimates $\underline{e}'\underline{y}$, $\underline{\ell}_1'\underline{y}$, $\underline{\ell}_2'\underline{y}$, $\ldots, \underline{\ell}_{(n-1)}'\underline{y}$ are independent, each with a single degree of freedom, and that together they provide the total sum of squares $\underline{y}'\underline{y}$. Since

$S.S.(\underline{e}'\underline{y}) = n\overline{y}^2$ and $S.S.(\underline{\ell}_i'\underline{y}) = (\underline{\ell}_i'\underline{y})^2/(\underline{\ell}_i'\underline{\ell}_i)$ and, also $\underline{\ell}_i'\underline{\ell}_i = 4/n$, it follows that

$$\underline{y}'\underline{y} - n\overline{y}^2 = (n/4) \sum_{i=1}^{n-1} (\underline{\ell}_i'\underline{y})^2 \qquad (2.1)$$

Suppose now that m observations are missing. We can now estimate only (n-m) effects from our (n-m) remaining observations, and we can choose which (n-m) of the original (n-1) effects, apart from the mean, are to be estimated and which m are to be sacrificed. Furthermore once the m sacrificed effects have been selected, we can regard, as the residual sum of squares to be minimized with respect to the m missing observations, that portion of the right hand side of (2.1) which arises from sacrificed effects. Minimization of this residual sum of squares will lead to m equations in the m missing values. However, in some cases the equations will be dependent and no solution will be possible. This difficulty is overcome by a careful choice of which effects are to be sacrificed. Once this choice is correctly made, the m equations which involve the missing values may be written down without any differentiation at all since every equation is obtained by setting equal to zero one of the sacrificed effects. This may be seen as follows. Suppose we decide to sacrifice the (n-m)-th to (n-1)-st effects and treat

$$S = (n/4) \sum_{i=n-m}^{n-1} (\underline{\ell}_i' \underline{y})^2 \qquad (2.2)$$

as the residual sum of squares. If we differentiate with respect to each of the m missing observations, put the result equal to zero, and drop out constant factors we shall obtain a series of m equations of the form

$$\sum_{i=n-m}^{n-1} a_{ij} \underline{\ell}_i' \underline{y} = 0 \qquad (j=1,2,\ldots,m) \qquad (2.3)$$

where a_{ij} is ± 1 according as the coefficient of the j^{th} missing observation in $\underline{\ell}_i' \underline{y}$ is positive or negative. Thus if the m by m matrix $\underline{A} = \{a_{ij}\}$ is non-singular (or, equivalently, the a_{ij} are such that equations (2.3) are independent) then equations (2.3) reduce to

$$\underline{\ell}_i'\underline{y} = 0, \quad i = (n-m),\ldots,(n-1). \qquad (2.4)$$

Correct choice of effects to be sacrificed therefore amounts to finding effects for which $\underline{A}$ is non-singular. This done, we shall have the m equations (2.4) in m unknowns. These could be solved for the missing value estimates but, in fact, solution of the equations is quite unnecessary, as the examples to be presented later will show.

So far our suggested procedure has involved setting equal to zero selected sacrificed effects. Hopefully the true value of these effects, if actually estimated, would be zero or close to zero. What if, in fact, we have set to zero one or more linear combinations of effects whose true value is far from zero? In such a case our procedure will usually cause the estimates of missing values to induce serious biases in the estimates which have been retained and consequently the missing value estimates would be unsatisfactory. Serious biases caused in this way can usually be detected by examining a half-normal plot of all the retained effects, after their estimation. As Daniel (1959) remarks, the small effects in a half-normal plot should, approximately, point towards the origin. If they do not, bias has been introduced, either by the missing value estimates or by another observation.

Strictly speaking, the estimated effects in a half-normal plot should be independent and this will no longer be true. The size of the correlations introduced by using estimates of missing values in a two level factorial or fractional factorial will depend principally on the number of observations missing and the number of runs in the whole design. These correlations will be small if the ratio (number of observations missing)/(number of runs) is small.

If we consider the estimation of factorial effects from a least squares viewpoint we know that when the design is complete, the variance-covariance matrix of the factorial effects is proportional to $(\underline{X}'\underline{X})^{-1} \sigma^2$ using a standard notation. This will be a diagonal matrix for a complete two level factorial and many fractional factorial designs. Let $\underline{X}_1$ be obtained from $\underline{X}$ by deleting all rows which correspond to missing observations and all columns which correspond to combinations of effects to be sacrificed. Then $(X_1'X_1)^{-1} \sigma^2$ will be the new variance-covariance natrix of the estimated combinations of effects. Consider a two level factorial or fractional factorial design for which $\underline{X}'\underline{X}$ is diagonal with elements n. If one observation is missing $\underline{X}_1'\underline{X}_1$ consists of diagonal elements (n-1) and off-diagonal elements ± 1. If two observations are missing $\underline{X}_1'\underline{X}_1$ consists of diagonal elements (n-2) and off-diagonal elements some of which are zero and some ± 2, and so on. By this heuristic approach we can see that if the number of missing values is only a small fraction of the number of runs in the design any correlations between effects which arise due to the use of missing value estimates are likely to be small and so will not greatly affect the use of the half-normal plot technique. Where the correlations are large the situation would be one in which we should not want to use a missing value technique in any case.

The estimation of missing values in the way described above and the subsequent analysis are both simple to carry out in practice. It is hoped that the examples which follow will demonstrate this.

3. One missing observation

(a) One missing value; full 2^k factorial design.

Choose an effect (main effect or interaction) which might be thought, a priori, to be equal to zero. This will normally be the highest order interaction between k factors though if experimental reasons dictated otherwise, another interaction might be set equal to zero. Alternative decisions might be suggested by the fact, for example, that two of the factors were blocks and no interaction between blocks could be expected.

Use of the highest order interaction in complete factorials is also suggested by Wright (1958), who discusses estimation of one missing value in a general factorial experiment and compares the results of using various sets of interactions as error for missing value purposes. Goulden (1952) gives this method for a 2^2 factional also. Treatment of the fractional factorial situation is not, however, mentioned.

Numerical example: A full 2^3 factorial design was performed on factors 1, 2 and 3 and the results, in standard order, were as in Table 1.

Table 1: Results from a 2^3 factorial design

1	2	3	12	13	23	123	y
-	-	-	+	+	+	-	10
+	-	-	-	-	+	+	16
-	+	-	-	+	-	+	2
+	+	-	+	-	-	-	22
-	-	+	+	-	-	+	8
+	-	+	-	+	-	-	20
-	+	+	-	-	+	-	2
+	+	+	+	+	+	+	44

The eighth result was observed but was completely out of line with the other results. An investigation showed that the test material changed its form at the extreme set of experimental conditions and so it was decided to substitute a missing value estimate m instead. The highest order interaction 123 was set equal to zero, i.e. $-10 + 16 + 2 - 22 + 8 - 20 - 2 + m = 0$ giving, upon solution, an estimate $m = 28$. The estimates of the other six effects, excluding the mean, were then computed using the missing value estimate $m = 28$ in place of the eighth observation and provided the values

1 effect = 16	12 effect = 7
2 effect = 0	13 effect = 3
3 effect = 2	23 effect = 1

A result heavily biased in this way is detected either by sight, experience, or by examining a half-normal plot of the effects calculated from the original data. As mentioned above a biased result will cause the smaller effects plotted to point, not at the origin but somewhere along the horizontal axis instead. Half-normal plots of estimates of effects before and after the missing value estimation are shown in Figure 1. It is clear that the original effects were biased and removal of the eighth observation, and its estimation, have removed this bias, since the

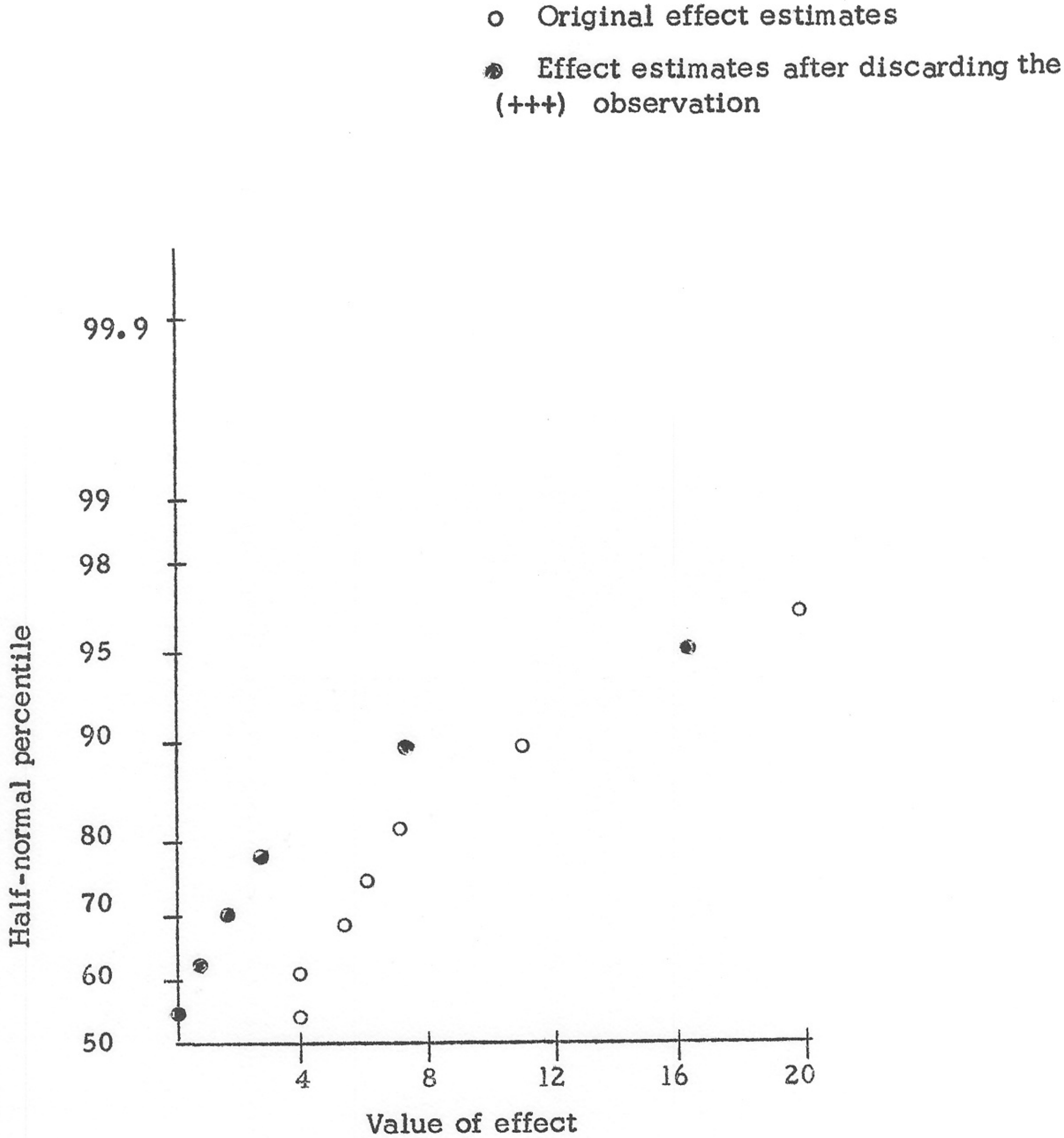


Figure 1. Half-normal plots of effects from a 2^3 factorial design before and after discarding the (+++) observation.

smaller effects point satisfactorily towards the origin.

(b) One missing value; fractional factorial design

In this situation the n independent linear combinations of the y's measure combinations of effects, not individual effects. Thus, if any combination of high order effects only is available it is an excellent candidate for setting equal to zero. For example a half fraction of a 2^6 generated by I = 123456 has 123 aliased with 456 so the combination of y's which estimates 123 + 456 is probably estimating a small effect and can be set equal to zero. Other $\underline{\ell}'\underline{y}$'s involve two factor interactions or main effects and so would not be used unless a priori reason dictated it.

In some experiments however the fractionation is so great that many or perhaps every linear combination of the y's estimates a main effect aliased with other effects. The procedure that should be used in such a case is as follows. Choose a "likely" estimate and set it to zero. Work out the other effects using the missing value estimated in the calculation and construct a half normal plot of them. If it does not point towards the origin, setting that particular combination of effects equal to zero has caused a bias in the results. Another combination should be chosen and the procedure repeated.

<u>Example</u>: Box and Hunter (1961) gave an example of the use of a 2^{7-4}_{III} design. The design and data are shown in Table 2.

Table 2: Results from a 2^{7-4}_{III} design

		1	2	3	4	5	6	7	Filtration Time
Experiment Number	1	-	-	-	-	+	+	+	68.4
	2	+	-	-	+	-	-	+	77.7
	3	-	+	-	+	-	+	-	(66.4)m
	4	+	+	-	-	+	-	-	81.0
	5	-	-	+	+	+	-	-	78.6
	6	+	-	+	-	-	+	-	41.2
	7	-	+	+	-	-	-	+	68.7
	8	+	+	+	+	+	+	+	38.7

The generators are 125, 136, 237, and 1234.

Denoting the estimates of effects by $\underline{e}'\underline{y}$, $\underline{\ell}_1'\underline{y}$, $\underline{\ell}_2'\underline{y}, \ldots, \underline{\ell}_7'\underline{y}$ and assuming that interactions between three or more variables can be ignored, then

$$\underline{e}'\underline{y} = \text{Mean}$$
$$\underline{\ell}_1'\underline{y} = 1 + 25 + 36 + 47$$
$$\underline{\ell}_2'\underline{y} = 2 + 15 + 37 + 46$$
$$\underline{\ell}_3'\underline{y} = 3 + 16 + 27 + 45$$
$$\underline{\ell}_4'\underline{y} = 4 + 35 + 26 + 17$$
$$\underline{\ell}_5'\underline{y} = 5 + 12 + 34 + 67$$
$$\underline{\ell}_6'\underline{y} = 6 + 13 + 24 + 57$$
$$\underline{\ell}_7'\underline{y} = 7 + 23 + 14 + 56$$

In many situations where designs such as these are used it is expected that only a few of the effects will be real ones. Thus although main effects are involved in all combinations of effects, not all will be appreciable. Thus some of the combinations of effects may very well be small.

Assume, now, that the third observation is missing. Denote it by m. Then the value of the seven combinations of effects are in terms of known results and m,

1 + 25 + 36 + 47 = 5.7- m/4;	when	zero,	m	= 22.8
2 + 15 + 37 + 46 =-19.4+ m/4;	"	"		= 77.6
3 + 16 + 27 + 45 = 0 - m/4;	"	"		= 0
4 + 35 + 26 + 17 =-16.1+ m/4;	"	"		= 64.4
5 + 12 + 34 + 67 = 19.8- m/4;	"	"		= 79.2
6 + 13 + 24 + 57 =-39.4+ m/4;	"	"		=157.6
7 + 23 + 14 + 56 = 13.2- m/4;	"	"		= 52.8

Setting any of these to zero provides an estimate of m but these vary from zero to 157.6. Which do we use? We can, for the purposes of this example, try all, one by one, and calculate the other six effects that can be estimated in each case. The results are shown in Table 3.

Table 3: Estimates of effects using the missing value estimate

Missing value estimate m	Effect combination** 1	2	3	4	5	6	7
22.8	*	-13.7	- 5.7	-10.4	14.1	-33.7	7.5
77.6	-13.7	*	-19.4	3.3	0.4	-20.0	- 6.2
0	5.7	-19.4	*	-16.1	19.8	-39.4	13.2
64.4	-10.4	- 3.3	-16.1	*	3.7	-23.3	- 2.9
79.2	-14.1	0.4	-19.8	3.7	*	-19.6	- 6.6
157.6	-33.7	20.0	-39.4	23.3	-19.6	*	-26.2
52.8	- 7.5	- 6.2	-13.2	- 2.9	6.6	-26.2	*

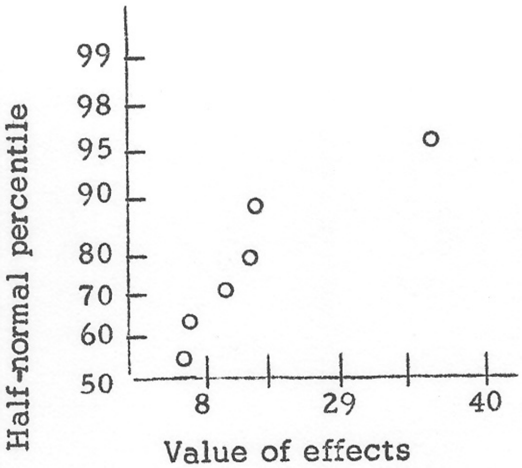
* Used for estimation

** Effect shown plus aliased effects

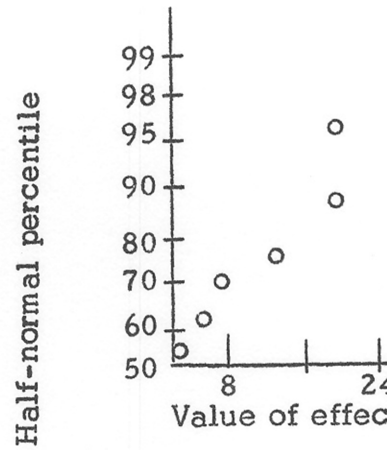
If now we plot the six effects which arise in each case, we obtain the seven half-normal plots in Figure 2. Some of the resulting plots exhibit very clearly considerable bias from an observation and the poor missing value estimate is the cause of the trouble in these cases. We see that estimates m = 77.6, 64.4, 79.2 and 52.8 are all in the range of reasonable estimates and that m = 22.8, 0, and 157.6 are not.

Note that any of the four "realistic" missing value estimates provide, as largest effects, those associated with main effects 1, 3 and 6, a tentative conclusion that was confirmed

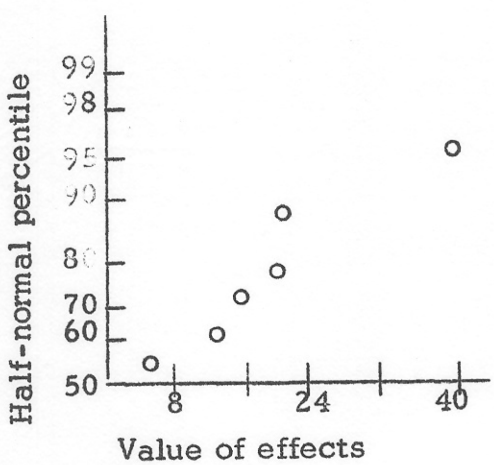
Figure 2



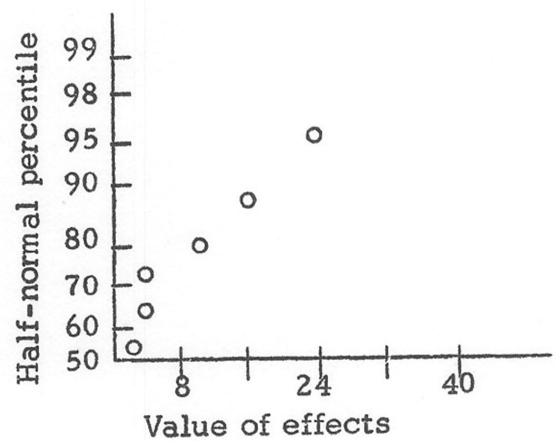
(1+25+34+ 47 discarded)



(2+15+37+46 discarded)



(3+16+27+45 discarded)



(4+35+26+17 discarded)

(continued)

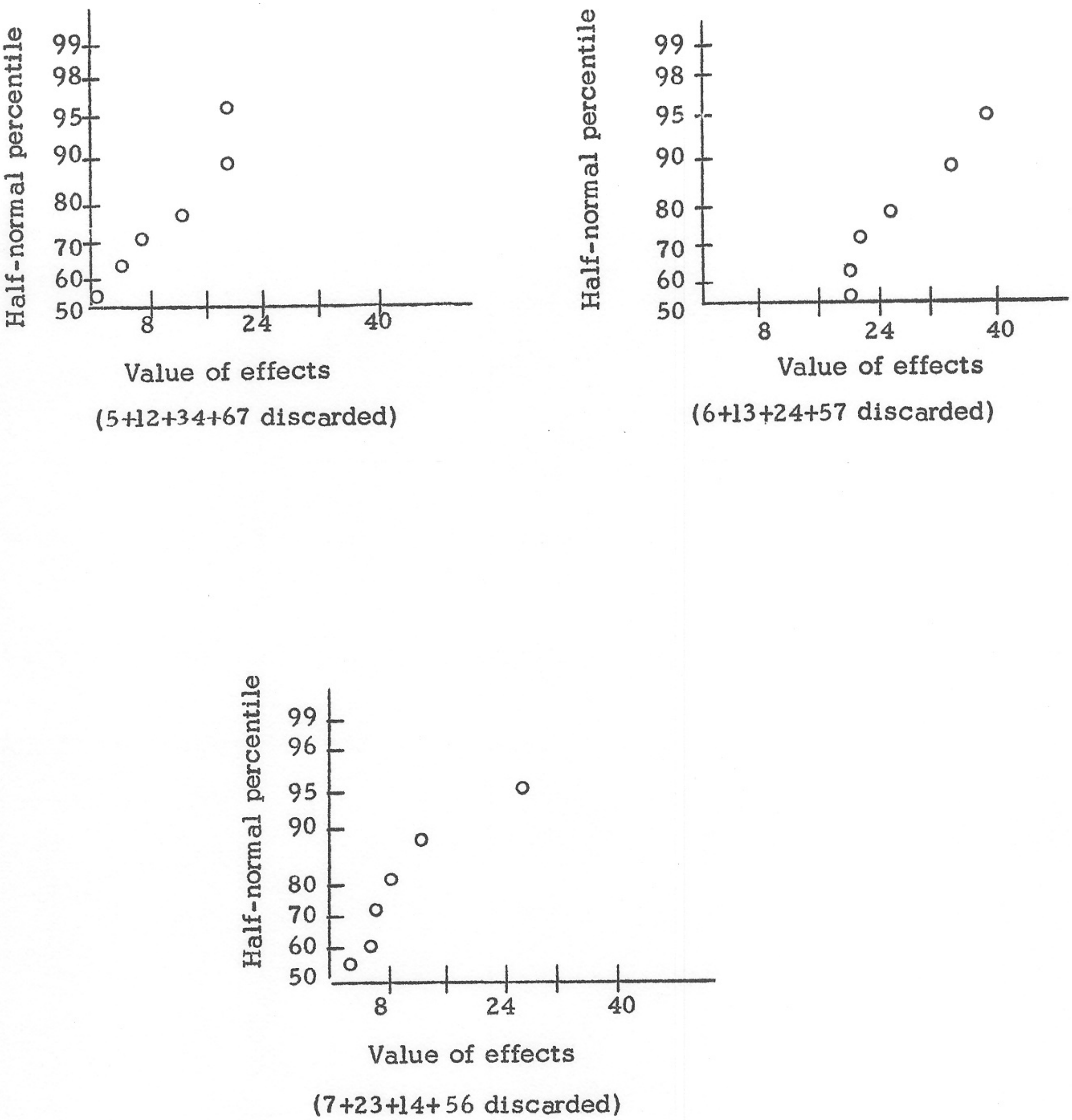


Figure 2. Half-normal plots of remaining effects in a 2_{III}^{7-4} design when one effect is sacrificed for missing value estimation. (Sacrificed effects are shown below each diagram).

when the actual experiment was continued in Box and Hunter (1961). In practice it should not be necessary to examine all possible missing value estimates. All that should be needed is to examine selected estimates until a satisfactory one is found. In the example above, the first of the values m = 77.6, 64.4, 79.2, and 52.8 which the experimenter tried could have been used even though Figure 2 shows that some choices are slightly better than others. Although the numerical magnitude of the effects would have been different in the four cases, the importance of the effects (1 + 25 + 36 + 47), (3 + 16 + 27 + 45) and (6 + 13 + 24 + 57) would have been recognized in every case.

If one observation is missing and another observation is badly biased as well it may happen that no missing value estimate seems satisfactory. In suwh a case one should attempt to find which other observation is causing the trouble and, if it can be found, remove it and treat that observation as "missing" too. It is usually fairly easy to spot such observations; see the discussion by Daniel (1959).

4. Two observations missing.

The data of Tables 1 and 2, previously used as examples, will be inadequate for a satisfactory illustration of the two missing values situation. From the practical point of view two observations missing out of eight in such a highly fractionated experiment would be a very unsatisfactory state of affairs, and missing value estimation in such circumstances would probably be used only if further experiments were impossible.

Consider now the following example. Although constructed, it is an adaption of an example given by Cochran and Cox (1957) on the texture of cake icing. A 2_{IV}^{6-2} fractional factorial design with defining relation $I = 1234 = 3456 = 1256$ is performed. The design, with observations, the second and seventh of which, y_2 and y_7, are missing is given in Table 4 as the first six columns. Other columns provide the combinations of plus and minus signs which are needed to obtain the combination of effects shown at the head of each column. For example, using the seventh column, an estimate of the effect combinations 12 + 34 + 56 would be provided by $(233 - y_2 - 221 + 317 + \ldots + 349)/8$ if y_2 and y_7 were known. We have assumed, in saying this, that we shall ignore all effects (such as 1235) which involve more than three factors. Such an assumption (or a stronger one) is frequently made in circumstances where fractional factorial designs are useful. For this example we use the first column of observations shown in Table 4.

It is useful as a first step in the estimation procedure to

Table 4. Results from a 2_{IV}^{6-2} design.

Effect Reference #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	y*	y**
Effect Combination	1=234 =256	2=134 =156	3=124 =456	4=123 =356	5=126 =346	6=125 =345	12=34 =56	13=24	14=23	15=26	16=25	35=46	36=45	135= 245= 236= 146	145= 235= 136= 246		
	-	-	-	-	-	-	+	+	+	+	+	+	+	-	-	263	263
	+	-	-	+	-	+	-	-	+	-	+	+	-	+	-	y_2	y_2
	-	+	-	+	-	+	-	+	-	+	-	+	-	-	+	251	251
	+	+	-	-	-	-	+	-	-	-	-	+	+	+	+	347	347
	-	-	+	+	-	-	+	-	-	+	+	-	-	+	+	385	385
	+	-	+	-	-	+	-	+	-	-	+	-	+	-	+	259	259
	-	+	+	-	-	+	-	-	+	+	-	-	+	+	-	y_7	237
	+	+	+	+	-	-	+	+	+	-	-	-	-	-	-	302	302
	-	-	-	-	+	+	+	+	+	-	-	-	-	+	+	155	155
	+	-	-	+	+	-	-	-	+	+	-	-	+	-	+	185	185
	-	+	-	+	+	-	-	+	-	-	+	-	+	+	-	135	y_{11}
	+	+	-	-	+	+	+	-	-	+	+	-	-	-	-	235	235
	-	-	+	+	+	+	+	-	-	-	-	+	+	-	-	401	401
	+	-	+	-	+	-	-	+	-	+	-	+	-	+	-	363	y_{14}
	-	+	+	-	+	-	-	-	+	-	+	+	-	-	+	347	347
	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	379	379

*Use this column for the example of 2 missing values. ** Use this column for the example of 3 missing values.

evaluate eight times the estimated effects with the missing observations included. Using the column reference numbers of Table 4, we find for eight times the effects:

$$
\begin{aligned}
1 &: \quad 133 + (y_2 - y_7) = L_1 \\
2 &: \; - 15 - (y_2 - y_7) = L_2 \\
3 &: \quad 865 - (y_2 - y_7) = L_3 \\
4 &: \quad 69 + (y_2 - y_7) = L_4 \\
5 &: \quad 393 - (y_2 + y_7) = L_5 \\
6 &: \; -647 + (y_2 + y_7) = L_6 \\
7 &: \quad 927 - (y_2 + y_7) = L_7 \\
8 &: \quad 207 - (y_2 + y_7) = L_8 \\
9 &: \; -749 + (y_2 + y_7) = L_9 \\
10 &: \quad 115 - (y_2 - y_7) = L_{10} \\
11 &: \; - 1 + (y_2 - y_7) = L_{11} \\
12 &: \quad 695 + (y_2 - y_7) = L_{12} \\
13 &: \; - 69 - (y_2 - y_7) = L_{13} \\
14 &: \; -479 + (y_2 + y_7) = L_{14} \\
15 &: \quad 609 - (y_2 + y_7) = L_{15}
\end{aligned}
\qquad (4.1)
$$

It will be noted that, in (4.1), effects with reference numbers 1, 2, 7, 12, 13, 14, 15 (call this "Group A") all involve $\pm(y_2 + y_7)$, while the remaining effects, with reference numbers 3, 4, 5, 6, 8, 9, 10, 11 (call this "Group B") all involve

$\pm(y_2 - y_7)$.

In order to estimate the missing values y_2 and y_7, we shall set two estimates of effects equal to zero as earlier described. This is the same as putting $L_i = L_j = 0 (i \neq j)$. However in order that the two estimates chosen shall provide two independent equations, we must choose two columns of Figure 4 so that the second and seventh rows of these columns provide a pattern like,

$$\begin{matrix} + & - \\ + & + \end{matrix}, \quad \begin{matrix} + & + \\ + & - \end{matrix}, \quad \begin{matrix} + & - \\ - & + \end{matrix}, \text{ or } \quad \begin{matrix} - & + \\ - & - \end{matrix},$$

(or with columns interchanged), and not like, for example,

$$\begin{matrix} + & - \\ + & - \end{matrix}, \quad \begin{matrix} - & - \\ + & + \end{matrix}, \quad \begin{matrix} + & - \\ - & + \end{matrix}, \text{ or } \quad \begin{matrix} + & + \\ - & - \end{matrix},$$

(or with columns interchanged). If we attach unities to the signs we get the $\underline{A}$ matrices previously described. The first set are all non-singular, the second set all singular. These patterns may easily be found by looking at the signs attached to y_2 and y_7 in pairs L_i and L_j chosen from (4.1) and it is then clear that in order to provide the right patterns we must choose L_i so that i belongs to Group A and L_j so that j belongs to Group B.

For suppose we do not; for example an immediate choice might seem to be to use the effect combinations (135 + 245 + 236 + 146) and (145 + 235 + 136 + 246) and set them to zero. This is

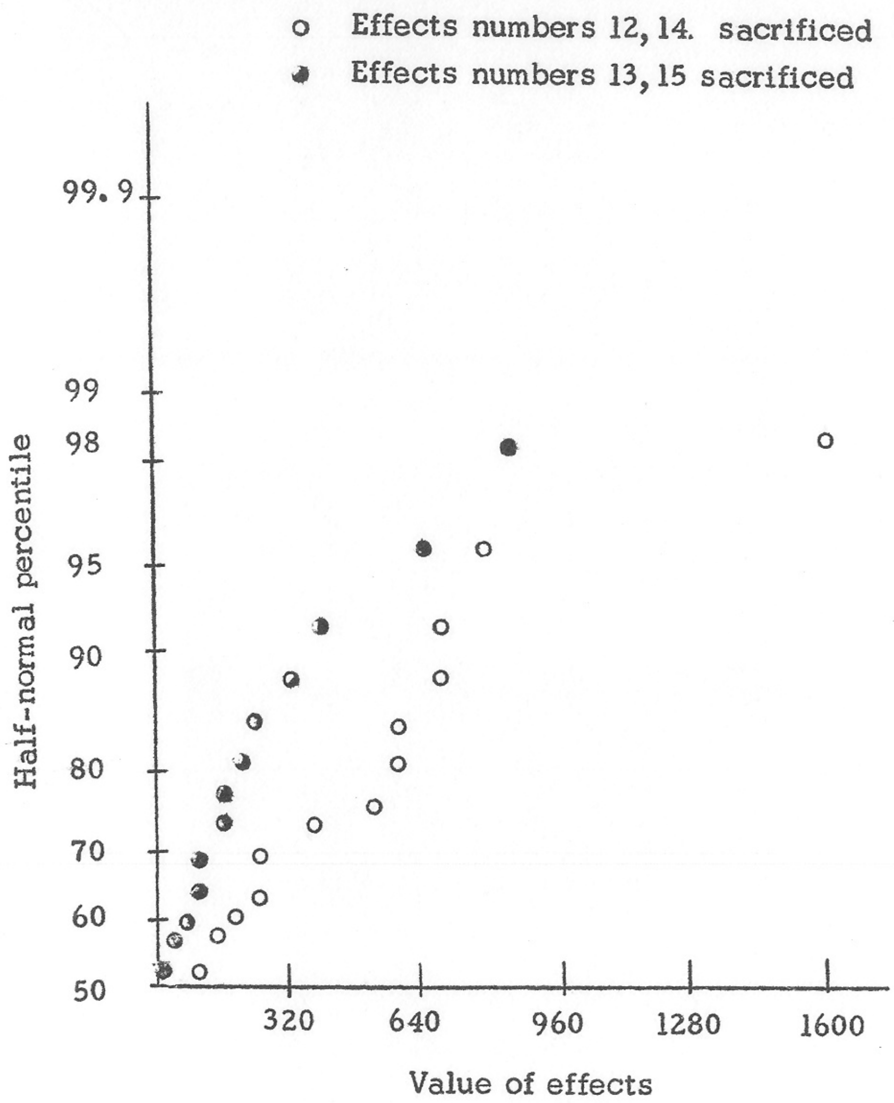


Figure 3. Plots of remaining effects after sacrificing two effects to estimate two missing values

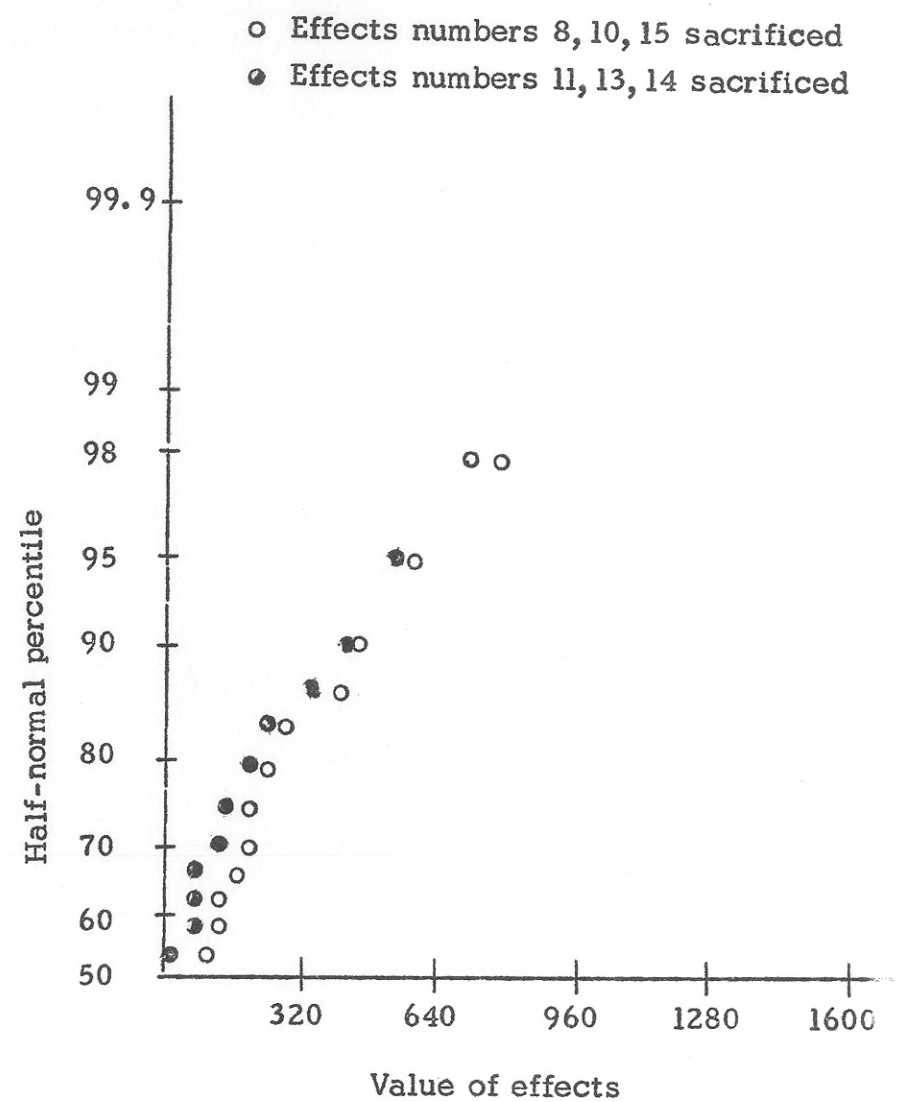


Figure 4. Plots of remaining effects after sacrificing three effects to estimate three missing values.

equivalent to choosing the residual sum of squares proportional to

$$S(14, 15) = (-y_1 + y_2 - y_3 + \ldots + y_{16})^2 + (-y_1 - y_2 + y_3 + \ldots + y_{16})^2$$
$$= L_{14}^2 + L_{15}^2,$$

say, where we are using the fourteenth and fifteenth column of Table 3. The estimation equations for the missing y_2 and y_7 would then be

$$\frac{\partial S}{\partial y_2} \equiv 2\ (L_{14} + L_{15}) = 0$$

$$\frac{\partial S}{\partial y_7} \equiv -2\ (L_{14} + L_{15}) = 0$$

which, clearly, provide only one equation. However, if we abandon estimates of the effect combinations (35 + 46) and (135 + 245 + 236 + 146) we shall obtain a residual proportional to $S(12, 14) = L_{12}^2 + L_{14}^2$, whence

$$\frac{\partial S}{\partial y_2} \equiv 2(L_{12} + L_{14}) = 0, \quad \text{and}$$

$$\frac{\partial S}{\partial y_7} \equiv -2(L_{12} - L_{14}) = 0$$

provide two independent equations which reduce to $L_{12} = L_{14} = 0$. It will be noticed that there are altogether 105 possible pairs of effect combinations. Only 56 of these pairs will give rise to two independent equations, and some of these latter 56 will induce biases in the remaining effects unless <u>all</u> effects are actually small. The 56 pairs which will provide two independent equations involve L_i and L_j where i is chosen from Group A and j from Group B.

Many of these pairs would not come into consideration, however. For example, it would usually be unwise to select initially a pair with a number between 1 and 6 inclusive since these numbers involve main effects. Normally one would choose a pair involving interactions of as high order as possible. Suppose we decided to choose the pair (12, 14). For estimation equations we find that

$$\begin{aligned} L_{12} = 0 &\text{ implies } y_7 - y_2 = 695, \\ L_{14} = 0 &\text{ implies } y_7 + y_2 = 479. \end{aligned} \qquad (4.2)$$

It is not necessary to solve these two equations in order to evaluate all the other 13 effect estimates since, as noted above, effects with reference numbers 1, 2, 7, 12, 13, 14, 15 ("Group A") all involve $\pm(y_2 + y_7)$ while the remaining effects, with reference numbers 3, 4, 5, 6, 8, 9, 10, 11 ("Group B"), all involve $\pm (y_2 - y_7)$. Thus the values in the right hand side of equations (4.2) can be immediately applied without solving the equations. Using this fact we obtain for the other 13 effects values which are one-eighth of the numbers given next to the effect combinations below.

1 + 234 + 256	: - 562
2 + 134 + 156	: 680
3 + 124 + 456	: 1560
4 + 123 + 356	: - 626
5 + 126 + 346	: - 86
6 + 125 + 345	: - 168
12 + 34 + 56	: 448
13 + 24	: - 272
14 + 23	: - 266
15 + 26	: 810
16 + 25	: - 696
35 + 46	: used for estimation
36 + 45	: 626
135 + 245 + 236 + 146	: used for estimation
145 + 235 + 136 + 246	: 130

A half-normal plot of these numerical values is shown in Figure 3. Since the smaller estimates plotted do not point at the origin, it appears that the estimates used here introduced considerable bias and thus the effects used for estimating y_2 and y_7 which were set equal to zero could not in fact have been small. Thus another choice must be made and the effects re-estimated. Let us try another pair, for example (13,15). The estimates are given by the two equations

$$L_{13} = 0 \quad \text{which implies} \quad y_2 - y_7 = -69$$

$$L_{15} = 0 \quad \text{which implies} \quad y_2 + y_7 = 609$$

Using these values, we find for eight times the remaining

13 estimates of combination of effects in the same order as before:

1 + 234 + 256	64
2 + 134 + 156	54
3 + 124 + 456	934
4 + 123 + 356	0
5 + 126 + 346	-216
6 + 125 + 345	- 38
12 + 34 + 56	318
13 + 24	-402
14 + 23	-136
15 + 26	184
16 + 25	- 70
35 + 46	626
36 + 45	used for estimation
135 + 245 + 236 + 146	130
145 + 235 + 136 + 246	used for estimation

A half-normal plot of these numerical values is shown in Figure 3. Here the smaller estimates plotted do point at the origin, and thus it appears that the estimates used for the missing values are performing their role satisfactorily. The analysis may now be continued in conventional ways.

5. Three observations missing

For purposes of illustration we shall again use the design of Table 4. Assume for an example that the second, eleventh and fourteenth observations are missing now, as shown in the second column of observations. Using the column reference numbers of Table 4 we find, for eight times the estimated effects:

$$
\begin{aligned}
1&: -332 + (y_2 - y_{11} + y_{14}) = L_1\\
2&: \ \ 450 - (y_2 - y_{11} + y_{14}) = L_2\\
3&: \ \ 874 - (y_2 + y_{11} - y_{14}) = L_3\\
4&: \ \ 60 + (y_2 + y_{11} - y_{14}) = L_4\\
5&: -342 - (y_2 - y_{11} - y_{14}) = L_5\\
6&: \ \ 88 + (y_2 - y_{11} - y_{14}) = L_6\\
7&: \ \ 1188 - (y_2 + y_{11} + y_{14}) = L_7\\
8&: -528 - (y_2 - y_{11} - y_{14}) = L_8\\
9&: -10 + (y_2 - y_{11} - y_{14}) = L_9\\
10&: \ \ 124 - (y_2 + y_{11} - y_{14}) = L_{10}\\
11&: -10 + (y_2 + y_{11} - y_{14}) = L_{11}\\
12&: \ \ 230 + (y_2 - y_{11} + y_{14}) = L_{12}\\
13&: \ \ 396 - (y_2 - y_{11} + y_{14}) = L_{13}\\
14&: -740 + (y_2 + y_{11} + y_{14}) = L_{14}\\
15&: \ \ 870 - (y_2 + y_{11} + y_{14}) = L_{15}
\end{aligned}
$$

In order to estimate the missing values y_2, y_{11}, y_{14}, we set three estimates of effect combinations equal to zero. In order to obtain three independent equations we must choose three L_i which are such that their coefficients of y_2, y_{11} and y_{14} provide a non-singular 3 by 3 matrix. This can be achieved, in fact, in 44 ways using only L_i for which $i = 7,8,\ldots,15$. (As before we might decide not to use L_i for which $i = 1,\ldots,6$ since these involve main effects.) This may be seen as follows. The

15 L_i above divide into four groups:

Group A: $i = 1, 2, 12, 13,$ involving $(y_2 - y_{11} + y_{14})$
Group B: $i = 3, 4, 10, 11,$ involving $(y_2 + y_{11} - y_{14})$
Group C: $i = 5, 6, 8, 9,$ involving $(y_2 - y_{11} - y_{14})$
Group D: $i = 7, 14, 15,$ involving $(y_2 + y_{11} + y_{14})$

It is clear that the choice of 3 L_i which give rise to a non-singular $\underline{A}$ matrix implies choosing an L from 3 distinct groups above. If we decide not to use an L_i involving a main effect ($i = 1,\ldots,6$) we are reduced to 2 choices from A, 2 from B, 2 from C and 3 from D. Since we can choose three groups in the four ways ABC, ABD, ACD or BCD, total choices involving $i = 7,\ldots,15$ reduce to $2^3 + 2^2.3 + 2^2.3 + 2^2.3 = 44$. As in previous examples we must make a choice, evaluate the remaining effects and plot them in a half-normal plot.

Suppose we choose 8, 10, and 15 from Groups C, B, and D respectively. The three estimation equations are then

$$L_8 = 0 \text{ which implies } y_2 - y_{11} - y_{14} = -528$$
$$L_{10} = 0 \text{ which implies } y_2 + y_{11} - y_{14} = 124$$
$$L_{15} = 0 \text{ which implies } y_2 + y_{11} + y_{14} = 870$$

Again, solution of equations such as these is unnecessary. For example since $L_8 = 0$ implies $y_2 - y_{11} - y_{14} = -528$ we can immediately obtain all effects of Group C. Similarly we can obtain the effects of Groups B and D by use of $L_{10} = 0$ and

$L_{15} = 0$. For Group A we can use the fact that $(y_2 - y_{11} + y_{14})$, which we need, can be obtained as 218 using the identity

$$(y_2 - y_{11} + y_{14}) + (y_2 + y_{11} - y_{14}) = (y_2 - y_{11} - y_{14}) + (y_2 + y_{11} + y_{14}) \qquad (5.1)$$

so that it is not necessary to know the individual estimates of the missing values at all. We find by this method, for eight times the estimates of the remaining 12 effect combinations, the following numerical values:

Reference No.	
1:	-114
2:	232
3:	750
4:	184
5:	186
6:	-440
7:	-318
8:	used for estimation
9:	-538
10:	used for estimation
11:	114
12:	448
13:	178
14:	130
15:	used for estimation

A half-normal plot of the effects is shown in Figure 4. We can see that the results exhibit some bias. Because of this we shall choose another set of 3 estimates of effect combinations and repeat.

Suppose we choose instead 11, 13, and 14 from Groups B, A and D respectively. The three estimation equations are then

$$\begin{aligned} L_{11} &= 0 \text{ which implies } y_2 + y_{11} - y_{14} = 10 \\ L_{13} &= 0 \text{ which implies } y_2 - y_{11} + y_{14} = 396 \\ L_{14} &= 0 \text{ which implies } y_2 + y_{11} + y_{14} = 740 \end{aligned}$$

These equations immediately give the estimates of effects in Groups B, A and D. For Group C, we find $(y_2 - y_{11} - y_{14}) = -334$, using the identity (5.1). Thus for eight times the estimates of the remaining 12 effect combinations, we find as follows.

Reference No.	1:	64
	2:	54
	3:	864
	4:	70
	5:	- 8
	6:	-246
	7:	448
	8:	-194
	9:	-344
	10:	114
	11:	used for estimation
	12:	626
	13:	used for estimation
	14:	used for estimation
	15:	130

A half-normal plot of the effects is shown in Figure 4. We see that the missing values have been satisfactorily estimated and the analysis may now be continued in conventional ways.

REFERENCES

Box, G. E. P. and Hunter, J. S. [1961a]. The 2^{k-p} fractional factorial designs. Technometrics 3, 311-351

Box, G. E. P. and Hunter, J. S. [1961b]. The 2^{k-p} fractional factorial designs. Technometrics 3, 449-458

Cochran, W. G. and Cox, G. M. [1957]. Experimental Designs. John Wiley and Sons, New York.

Daniel, C. [1959]. Use of half-normal plots in interpreting factorial two level experiments. Technometrics 1, 311-341.

Goulden, C. H. [1952]. Methods of Statistical Analysis. John Wiley and Sons, New York.

Wright, G. M. [1958]. The estimation of missing values in factorial experiments. New Zealand Journal of Science 1, 1-8