Retour au menu

LE PROBLEME DES DONNEES MANQUANTES

par D. A. PREECE
University of Kent - Grande Bretagne

RESUME : De temps en temps, on trouve des « données manquantes » parmi les résultats d'une expérience ; à cause d'un accident quelconque, une variable peut manquer de valeurs correspondantes à quelques-unes des parcelles. On peut analyser une telle variable incomplète par la méthode des moindres carrés, mais, en général, une telle analyse sera non orthogonale et la forme de l'analyse sera différents de celle projetée au commencement de l'expérience. Voilà pourquoi on a étudié plusieurs méthodes pour « estimer » les données manquantes. Ces méthodes comprennent des méthodes algébriques et des méthodes d'ordinateur. Pendant l'exposé on va discuter plusieurs de ces méthodes ; toutes les méthodes donnent les mêmes estimations des données manquantes. Après avoir estimé les données manquantes, on peut compléter l'analyse sans difficulté.

1. INTRODUCTION GENERALE.

Je suis enchanté d'avoir été invité à participer à ce séminaire. Quoique l'Université de Kent, où je travaille, n'ait pas de département d'agronomie, je m'intéresse beaucoup aux applications statistiques à l'agriculture. Avant d'établir ma demeure à Cantorbéry, j'ai travaillé pendant six ans dans le département de statistiques à Rothansted Experimental Station, département fondé par R. A. FISHER. A Rothamsted je me suis occupé surtout de la planification et de l'analyse des expériences agricoles ; j'ai travaillé sous la direction du docteur F. YATES. Puis, en 1969, je me suis rendu au comté de Kent, qui est célèbre pour sa récolte de fruits ; on y cultive surtout les pommes, les poires, les prunes, et les prunes de Damas. Dans le comté de Kent se trouve East Malling Research Station, station de recherches sur les problèmes de culture des arbres fruitiers. Le chef du département de statistiques à East Malling est mon collègue, le decteur S. C. PEARCE. Le docteur PEARCE et moi, nous nous intéressons tous les deux aux deux sujets que je vais discuter aujourd'hui.

Mon premier thème est « Le problème des données manquantes ». Je regrette beaucoup que je dois commencer en m'excusant de ma connaissance incomplète de la langue française ; je vous demande de bien vouloir me pardonner les erreurs linguistiques dans mon exposé.

2. LES VARIABLES AVEC DES DONNEES MANQUANTES.

Si l'une des parcelles d'un essai de champ est envahie par des animaux qui endommagent les plantes, la récolte de cette parcelle sera réduite, et la vraie récolte de la parcelle sera inconnue. On aura ainsi une « valeur manquante ».

Quelquefois on s'intéresse à plusieurs variables obtenues dans une seule expérience. (Ici le mot « variable » signifie un ensemble de valeurs numériques, une valeur par parcelle.) Dans un tel cas, il se peut bien que le nombre de valeurs manquantes soit différent pour les diverses variables ; et même si chaque variable a le même nombre de valeurs manquantes, les ensembles correspondants de parcelles peuvent être différents. Pourtant, les analyses dont je vais m'occuper aujourd'hui sont les analyses d'une seule variable ; je ne vais pas discuter les problèmes qu'on rencontre dans une analyse qui tient compte de plusieurs variables avec différentes parcelles manquantes.

Il arrive quelquefois qu'il y ait tant de données manquantes dans une seule variable qu'on ne peut pas estimer tous les paramètres du modèle linéaire de l'expérience qu'on a planifiée. Par exemple, si toutes les valeurs relatives à un certain traitement sont manquantes, la variable ne peut nous donner aucun renseignement sur le traitement, et on ne peut pas estimer un paramètre correspondant à ce traitement. Dans ce cas assez simple, on peut analyser les données en changeant le modèle, spécifiquement en omettant le paramètre correspondant au traitement manquant. Cependant, je vais supposer désormais que la distribution des valeurs manquantes soit telle qu'on peut encore estimer tous les paramètres du modèle premier.

Il est important de nous rappeler qu'une valeur de « zéro » n'est pas nécessairement une valeur manquante. Quelquefois dans les expériences sur les animaux, notamment dans les expériences où chaque animal constitue une parcelle, il arrive qu'un animal meure avant la fin de l'expérience. Si les raisons pour la mort de l'animal sont indépendantes des traitements de l'expérience, il s'agit peut-être d'une valeur manquante. Au contraire, si la mort résulte du traitement auquel l'animal a été exposé, nous ne devons pas supposer qu'il s'agisse d'une valeur manquante.

3. DEUX EXEMPLES.

Pour illustrer les méthodes pour traiter des données manquantes, je me servirai de deux exemples. Les données de ces exemples sont artificielles ; je les ai choisies de sorte que les calculs nécessaires soient faciles. Le plan de chaque exemple est en blocs aléatoires complets ; les valeurs manquantes sont indiquées par la lettre « x » :

Exemple 1

	Traitements		
Blocs	A	B	C
I	15	14	13
II	13	9	11
III	15	9	12
IV	13	8	x

EXEMPLE 2

Blocs	Traitements			
	A	*B*	*C*	*D*
I	x	904	901	867
II	752	815	x	913
III	878	947	997	954
IV	850	886	998	994

4. L'ANALYSE DE LA VARIABLE INCOMPLETE PAR LA METHODE DES MOINDRES CARRES.

La méthode la plus évidente pour analyser une variable qui a des valeurs manquantes est d'analyser la variable incomplète par la méthode des moindres carrés. Pour utiliser cette méthode, on ne tient aucun compte des parcelles avec les valeurs manquantes; on obtient les solutions des équations normales pour les données non manquantes. La forme d'une telle analyse est différente de celle projetée au commencement de l'expérience. D'ordinaire, une telle analyse est non orthogonale. Ceci peut ne pas être désavantageux, si l'on a un programme convenable sur ordinateur. Notamment, pour analyser les deux exemples ci-dessus, on peut se servir d'un programme général qui utilise la méthode de TOCHER (1952) — méthode d'analyse d'aucune expérience en blocs incomplets. Sans un tel programme général, cependant, l'analyse de la variable incomplète peut être très difficile. Voilà pourquoi on a développé des méthodes spéciales pour traiter des données manquantes. L'idée fondamentale de ces méthodes est d'estimer les données manquantes, c'est-à-dire de trouver des valeurs numériques qu'on insère dans la table de données à la place des données manquantes. Après avoir inséré les estimations, on complète l'analyse des données presque comme si aucune valeur n'était manquante.

5. LES PROCEDES POUR ESTIMER LES DONNEES MANQUANTES.

Tous les procédés pour estimer les données manquantes sont équivalents, car ils nous donnent tous les mêmes estimations. De plus, la somme des carrés de l'erreur dans l'analyse de variance des données augmentées est égale à la somme des carrés de l'erreur dans l'analyse de variance de la variable incomplète.

Parmi tous les procédés équivalents se trouvent des méthodes algébriques et des méthodes d'ordinateur. Parmi les méthodes d'ordinateur se trouvent des méthodes itératives et des méthodes non itératives. Je ne vais pas décrire toutes les méthodes; elles sont trop nombreuses, et quelques-unes d'entre elles

Retour au menu

ont été remplacées par des méthodes améliorées. Je ne vais décrire que les méthodes les plus importantes et les plus populaires.

6. METHODES ALGEBRIQUES.

6.1. L'INSERTION DES VALEURS QUI SONT « EGALES A LEURS ESPERANCES ESTIMEES » (DE LURY, 1946; WILKINSON, 1958).

Pour la première méthode algébrique, les estimations des données manquantes sont les valeurs qui sont « égales à leurs espérances estimées », c'est-à-dir que les estimations sont choisies de sorte que les résidus correspondants soient zéro.

La recette de cette méthode est la suivante. Soient x, y, z, ... les valeurs à estimer. Calculez les résidus pour les parcelles manquantes; ces résidus sont des fonctions de $x, y, z, \ldots$. Égalez à zéro chacun de ces résidus, et résolvez les équations ainsi obtenues pour trouver les estimations voulues.

Dans notre exemple numéro 1, nous avons

Blocs	Traitements *A*	*B*	*C*	Total
I	15	14	13	42
II	13	9	11	33
III	15	9	12	36
IV	13	8	x	$21 + x$
Total	56	40	$36 + x$	$132 + x$

Le résidu pour la parcelle manquante

$$= x - \frac{1}{3}(21 + x) - \frac{1}{4}(36 + x) + \frac{1}{12}(132 + x)$$

$$= \frac{1}{2}x - 5\,.$$

Lorsque nous égalons ce résidu à zéro, nous obtenons l'estimation $x = 10$.

Dans notre exemple numéro 2, nous avons

Blocs	Traitements *A*	*B*	*C*	*D*	Total
I	x	904	901	867	$2672 + x$
II	752	815	y	913	$2480 + y$
III	878	947	997	954	3776
IV	850	886	998	994	3728
Total	$2480 + x$	3552	$2896 + y$	3728	$12656 + x + y$

Les résidus pour les parcelles manquantes sont

$$x-\frac{1}{4}(2\,672+x)-\frac{1}{4}(2\,480+x)+\frac{1}{16}(12\,656+x+y)=(9x+y-7\,952)/16$$

et

$$y-\frac{1}{4}(2\,480+y)-\frac{1}{4}(2\,896+y)+\frac{1}{16}(12\,656+x+y)=(x+9y-8\,848)/16\,.$$

Nous devons ainsi résoudre les équations

$$9x+y-7952=0$$
$$x+9y-8848=0\,.$$

Il en suit que les estimations sont $x=784$ et $y=896$.

On peut se servir de cette méthode algébrique pour obtenir la formule de YATES (1933) pour l'estimation d'une seule valeur manquante dans un plan en blocs aléatoires complets. En commençant avec la table

Blocs	Traitements 1 2 3 … q	Total
1		
2		
3	x	$P+x$
⋮		
p		
Total	$Q+x$	$G+x$

on voit que le résidu pour la parcelle manquante est

$$x-\frac{1}{q}(P+x)-\frac{1}{p}(Q+x)+\frac{1}{pq}(G+x)=\frac{(p-1)(q-1)x-(pP+qQ-G)}{pq}$$

Il en suit que l'estimation est

$$x=\frac{pP+qQ-G}{(p-1)(q-1)}\,.$$

6.2. MINIMISATION DE LA SOMME DES CARRES DE L'ERREUR (YATES, 1933).

La deuxième méthode algébrique que je vais décrire est la méthode de R. A. FISHER, décrite par YATES en 1933. La recette est la suivante. Soient x, y,

Retour au menu

z, ... les valeurs à estimer. Calculez la somme des carrés de l'erreur comme fonction de $x, y, z, \ldots$. Choisissez les valeurs de x, y, z, ... qui minimisent la somme des carrés de l'erreur.

Dans notre exemple numéro 1, les sommes des carrés sont les suivantes:

Blocs	$78 - 8x + \frac{1}{4}x^2$
Traitements	$56 - 4x + \frac{1}{6}x^2$
Erreur	$58 - 10x + \frac{1}{2}x^2$
Total	$192 - 22x + \frac{11}{12}x^2$

Puisque

$$\frac{d}{dx}\text{(S.C. de l'erreur)} = -10 + x\,,$$

nous avons l'estimation $x = 10$ de nouveau.

Bien entendu, la méthode de FISHER est plus difficile que la méthode de DE LURY et de WILKINSON, mais celle de FISHER est très importante pour des raisons théoriques.

7. METHODES D'ORDINATEUR.

7.1. LA METHODE DE LA COVARIANCE (BARTLETT, 1937; COONS, 1957).

Parmi les méthodes d'ordinateur pour estimer les données manquantes, la méthode de la covariance est peut-être la mieux connue. Dans cette méthode, la variable principale y se compose des valeurs observées, avec un zéro pour chaque parcelle manquante. Il y a une variable auxiliaire pour chaque parcelle manquante; chacune de ces variables $x_1, x_2, \ldots$ ne se compose que de valeurs égales à zéro excepté que, dans chaque variable, la valeur pour la parcelle manquante correspondante est -1. Le modèle de l'analyse de la covariance peut s'écrire

$$y = \text{(termes du modèle correspondant de l'analyse de variance)} + \\ + c_1x_1 + c_2x_2 + \ldots$$

Les estimations des données manquantes sont $\hat{c}_1, \hat{c}_2, \ldots,$ où $\hat{c}_i$ est l'estimation

Retour au menu

habituelle de c_i. Notre analyse ultime est ainsi une analyse des valeurs « corrigées » de la variable y.

Pour notre exemple numéro 1, les deux variables, avec les totaux, sont les suivantes :

	y			
Blocs	A	B	C	Total
I	15	14	13	42
II	13	9	11	33
III	15	9	12	36
IV	13	8	0	21
Total	56	40	36	132

	x			
Blocs	A	B	C	Total
I	0	0	0	0
II	0	0	0	0
III	0	0	0	0
IV	0	0	-1	-1
Total	0	0	-1	-1

Il en suit que les sommes des carrés et des produits sont

	yy	xy	xx
Blocs	78	4	$\frac{1}{4}$
Traitements	56	2	$\frac{1}{6}$
Erreur	58	$E_{xy} = 5$	$E_{xx} = \frac{1}{2}$
Total	192	11	$\frac{11}{12}$

Il en suit que

$$\hat{c} = E_{xx}/E_{xy} = 5/(1/2) = 10\,.$$

7.2. La méthode de Rubin (Rubin, 1972).

La méthode de Rubin, qui a paru en 1972, est fondamentalement la même que la méthode de la covariance. La différence est que celle-ci n'emploie pas les matrices, tandis que Rubin calcule l'inverse d'une matrice qui a autant le lignes (ou colonnes) qu'il y a de données manquantes. Selon la méthode de Rubin, les estimations des valeurs manquantes sont les éléments du vecteur ligne $-\boldsymbol{\rho}\boldsymbol{R}^{-1}$ où

(a) le k-ième élément de $\boldsymbol{\rho}$ est le résidu obtenu pour la k-ième parcelle manquante lorsqu'on donne la valeur 0 à toutes les parcelles manquantes, et

Retour au menu

(*b*) la k-ième ligne de $\boldsymbol{R}$ se compose des résidus obtenus pour les parcelles manquantes, lorsqu'on donne la valeure 1 à la K-ième parcelle manquante et la valeur 0 à toutes les autres parcelles (manquantes ou non manquantes).

Dans notre exemple numéro 2,

$$\boldsymbol{\rho} = (-\,497 \qquad -\,553)$$

$$\boldsymbol{R} = \begin{pmatrix} 9/16 & 1/16 \\ 1/16 & 9/16 \end{pmatrix}$$

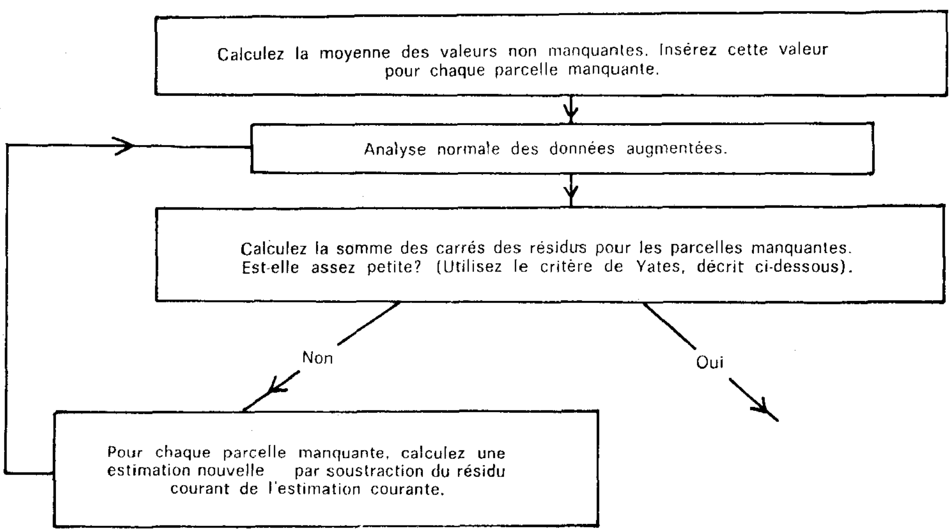
et

$$-\,\boldsymbol{\rho}\boldsymbol{R}^{-1} = (784 \qquad 896)\,.$$

Il y à des circonstances où la méthode de RUBIN est plus rapide que toutes les autres. Et vraiment on a dit que la méthode de RUBIN rend surannées les méthodes itératives que je vais décrire maintenant.

7.3. LE PROCESSUS ITERATIF DE HEALY ET WESTMACOTT (HEALY et WESTMACOTT, 1956; STAT-PACK, section 8, p. 27).

Le processus itératif le plus simple est celui de HEALY et WESTMACOTT. Les détails de ce processus, dit « The HEALY-WESTMACOTT process », étaient publiés en 1956, lorsque les auteurs travaillaient à Rothamsted. Je vais décrire ce processus à l'aide d'un diagramme:



Pour employer le critère de YATES, on calcule, à chaque itération, la rapport

$$X' = \frac{(\text{S.C. des résidus pour les parcelles manquantes})/n'}{(\text{S.C. de tous les residus à la 1ère itération})/E}$$

Retour au menu

où

n' = nombre de parcelles manquantes

et

E = nombre de d.l. de l'erreur, au cas où nulle parcelle n'est manquante.

Lorsque $X' < 0{,}005$, on peut arrêter le processus.

Dans notre exemple numéro 1, la moyenne des valeurs non manquantes est 12. La première estimation approximative pour la parcelle manquante est ainsi 12; le résidu correspondant est 1. Par conséquent, la 2ème approximation est $12 - 1 = 11$; le résidu correspondant est 1/2. En résumé, nous avons:

Itération	Estimation	Résidu	X'
1	12	1	0,600
2	11	$\frac{1}{2}$	0,150
3	$10\frac{1}{2}$	$\frac{1}{4}$	0,038
4	$10\frac{1}{4}$	$\frac{1}{8}$	0,009
5	$10\frac{1}{8}$	$\frac{1}{16}$	0,002

Le processus de HEALY et WESTMACOTT converge toujours.

7.4. LE PROCESSUS AMELIORE DE HEALY ET WESTMACOTT (PEARCE, 1965; PREECE, 1971).

Un processus qui est habituellement plus rapide que celui de HEALY et WESTMACOTT a été décrit par PEARCE en 1965 et par moi-même en 1971. La différence entre les deux methodes itératives est très simple: à chaque itération, au lieu de soustraire le résidu courant, on soustrait n/E fois le résidu courant, où n = le nombre de parcelles au cas où nulle d'entre elles est manquante.

Dans notre exemple numéro 1, $n = 12$ et $E = 6$, et nous avons $n/E = 2$ Par conséquent nous avons

Itération	Estimation	Résidu	X'
1	12	1	0,600
2	10	0	0,000

et nulle itération additionnelle n'est exigée.

Retour au menu

Dans la plupart des analyses factorielles avec une seule valeur manquante, la 2ème analyse du processus amelioré nous donne pour la parcelle manquante un résidu égal à zéro, afin que nulle itération additionnelle ne soit exigée. Pour des autres analyses, aussi, le processus amelioré est plus rapide que l'original.

8. L'ACHEVEMENT DE L'ANALYSE DE LA VARIANCE.

Après avoir estimé les données manquantes, il est facile de compléter une analyse approximative de la variance. Pour notre exemple numéro 1, où l'estimation de la valeur manquante est 10, l'analyse approximative est la suivante:

S.V.	d.l.	S.C.	C.M.
Blocs	3	23	$7\frac{2}{3}$
Traitements	2	$32\frac{2}{3}$	$16\frac{1}{3}$
Erreur	5	8	$1\frac{3}{5}$
Total	10	$63\frac{2}{3}$	—

Vous remarquerez que le nombre total de d.l. est 10 à cause de la parcelle manquante, et que, par conséquent, le nombre de d.l. de l'erreur est 5.

Comme je l'ai déjà dit, la somme des carrés de l'erreur dans l'analyse ci-dessus est la même que la somme des carrés de l'erreur dans l'analyse de variance de la variable incomplète. On peut ainsi dire que, dans notre analyse approximative, le carré moyen de l'erreur est *exact*. Mais le carré moyen des traitements n'est pas exact.

Pour trouver la S.C. exacte des traitments, nous devons estimer de nouveau la valeur manquante, sans tenir aucun compte des traitements. Soit x' l'estimation nouvelle. Puis le résidu pour la parcelle manquante est

$$x' - \frac{1}{3}(21 + x') .$$

Lorsque nous égalons ce résidu à zéro, nous voyons que l'estimation nouvelle est $x' = 10\frac{1}{2}$. Cette valeur nous donne la table

S.V.	d.l.	S.C.
Blocs (traitements ignorés)	3	$21\frac{9}{16}$
Résiduel	7	$40\frac{1}{2}$
Total	10	$62\frac{1}{16}$

Il en suit que la valeur exacte de la S.C. des traitements (blocs elimines)

$$\begin{aligned} &= \text{S.C. résiduel (ci-dessus)} - \text{S.C. de l'erreur (exacte)} \\ &= 40\tfrac{1}{2} - 8 \\ &= 32\tfrac{1}{2}. \end{aligned}$$

On peut obtenir la même valeur par la méthode de TOCHER.

Au premier abord, il y a des difficultés relatives aux moyens pour les traitements. Les moyens calculés après avoir estimé les valeurs manquantes ne sont pas en général les mêmes que ceux obtenus dans une analyse de la variable incomplète. Par exemple, la méthode de TOCHER nous donne les moyens

$$14\frac{3}{22} \quad 10\frac{3}{22} \quad 11\frac{7}{11}$$

pour les trois traitements de notre exemple numéro 1, alors que les méthodes qui estiment la valeur manquante nous donnent les moyens

$$14 \quad 10 \quad 11\frac{1}{2}.$$

Mais les différences entre les moyens sont les mêmes dans les deux cas et, au bout du compte, on fait une expérience pour *comparer* les traitements et pour évaluer les *différences* entre les traitements.

9. CONCLUSION.

En guise de conclusion, je dois vous dire que j'ai dû omettre, aujourd'hui, beaucoup de details tant théoriques que pratiques. Si vous voulez retrouver ces details, vous pourrez vous replonger dans la bibliographie que j'ai fournie

Retour au menu

en fin de mon texte. Dans la bibliographie j'ai inclu plusieurs articles non cités dans mon texte, y compris des articles au sujet du problème des valeurs « mixed-up », c'est-à-dire le problème où, à cause d'un accident, on ignore les valeurs individuelles pour un ensemble des parcelles, mais on connait la *somme* des valeurs inconnues. Les méthodes que j'ai decrites pour les données manquantes peuvent être modifiées pour les valeurs « mixed-up ».

10. EXEMPLE SUPPLEMENTAIRE.

(« Experimental Designs » par W. C. Cochran et Gertrude M. Cox, 2ème éd., pp. 110-112, Wiley).

Un plan en trois blocs aléatoires complets. L'expérience a testé cinq niveaux de potasse sur le cotonnier. On a calculé un « indice de puissance » pour un échantillon du coton de chaque parcelle, sauf que deux échantillons ont été endommagés avant leur arrivée au laboratoire:

	Niveaux de potasse				
Blocs	36	54	72	108	144
1	x	8,14	7,76	7,17	7,46
2	8,00	8,15	x	7,57	7,68
3	7,93	7,87	7,74	7,80	7,21

BIBLIOGRAPHIE

Bartlett M. S., *Some examples of statistical methods of research in agriculture and applied botany*, J. R. Statist. Soc. Suppl., **4**, 137-170 (1937).

Bose S. S., *The estimation of mixed-up yields and their standard errors*, Sankhya, **4**, 112-120 (1938).

Bose S. S. - Mahalanobis P. C., *On estimating individual yields in the case of mixed-up yields of two or more plots in field experiment*, Sankhya, **4**, 103-111 (1938).

Coons I., *The analysis of covariance as a missing plot technique*, Biometrics, **13**, 387-405 (1957).

De Lury D. B., *The analysis of Latin squares when some observations are missing*, J. Amer. Statist. Assoc., **41**, 370-389 (1946).

Draper N. R., *Missing values in response surface designs*, Technometrics, **3**, 389-398 (1961).

Hartley H. O., *A plan for programming analysis of variance for general purpose computers*, Biometrics, **12**, 110-122 (1956).

Healy M. J. R. - Westmacott M. H., *Missing values in experiments analysed on automatic computers*, Appl. Statist., **5**, 203-206 (1956).

Nair K. R., *The application of the technique of analysis of covariance to field experiments with several missing or mixed-up plots*, Sankhya, **4**, 581-588 (1940).

PEARCE S. C., *Biological Statistics: an Introduction*, p. 111 (1956).

PEARCE S. C. - JEFFERS J. N. R., *Block designe and missing data*, J. R. Statist. Soc. B, **33**, 131-136 (1971).

PREECE D. A., *Iterative procedures for missing values in experiments*, Technometrics, **13**, 743-753 (1971).

PREECE D. A. - GOWER J. C., *An iterative computer procedure for mixed-up values in experiments*, Appl.Statist. **23**, (1974).

RUBIN D. B., *A non-iterative algorithm for least squares estimation of missing values in any analysis of variance design*, Appl. Statist., **21**, 136-141 (1972).

SCLOVE S. T., *On missing value estimation in experimental design models*, Amer. Statistician, **26**, 2, 25-26 (1972).

SHEARER P. R., *Missing data in quantitative designe*, Appl. Statist., **22**, 135-140 (1973).

STAT-PACK, section 8, p. 27.

TOCHER K.D., *The design and analysis of block experiments*, J. R. Statist. Soc. B, **14**, 45-100 (1952).

WILKINSON G. N., *The analysis of covariance with incomplete data*, Biometrics, **13**, 363-372 (1957).

WILKINSON G. N., *Estimation of missing values for the analysis of incomplete data*, Biometrics, **19**, 257-286 (1958*a*).

WILKINSON G. N., *The analysis of variance and derivation of standard errors for incomplete data*, Biometrics, 14, 360-384 (1958*b*).

YATES F., *The analysis of replicated experiments when the field results are incomplete*, Empire J. Exp. Agric., **1**, 129-142 (1933); réimprimé dans: Experimental Design: Selected Papers, pp. 41-56.