Biometrika (1988), 75, 1, *pp.* 75-80
Printed in Great Britain

Optimal design for the estimation of variance components

BY RAHUL MUKERJEE

Stat.-Math. Division, Indian Statistical Institute, Calcutta 700 035, India

AND S. HUDA

Department of Statistics, King Saud University, Riyadh, Saudi Arabia

SUMMARY

The design problem for the estimation of variance components by the method of unweighted squares of means, under a multifactor random effects model, is considered. First it is shown that with the numbers of levels of the factors fixed, a balanced design, if it exists, is optimal. Next the numbers of levels are also treated as decision variables and the derivation of minimax designs is indicated.

Some key words: Balanced design; Factorial calculus; Holder's inequality; Random effects.

1. INTRODUCTION

The development of a systematic optimal design theory for estimating variance components has received attention in recent years. For previous work see Mostafa (1967), Anderson (1975, 1981), Patterson & Thompson (1971), Thompson & Anderson (1975), Muse & Anderson (1978), Ahrens & Pincus (1981) and also a recent review by Khuri & Sahai (1985). The present paper employs a factorial calculus (Kurkjian & Zelen, 1963) to derive results in an m-factor setting.

Consider a random effects model involving m factors $F_1, \ldots, F_m$, for the jth factor a random selection of s_j levels $(1 \leqslant j \leqslant m)$ being included in the experiment. A typical 'cell' is then $i = (i_1, \ldots, i_m)$ and let n_i be the number of observations in the ith cell $(0 \leqslant i_j \leqslant s_j - 1, 1 \leqslant j \leqslant m)$. The cells are assumed to be lexicographically ordered. A typical observation, Y_{iu}, the uth observation in the ith cell, is represented by the model

$$Y_{iu} = \mu + a_{i_1}^{(1)} + \ldots + a_{i_m}^{(m)} + a_{i_1 i_2}^{(12)} + \ldots + a_{i_{m-1} i_m}^{(m-1,m)} + \ldots + a_{i_1 i_2 \ldots i_m}^{(12\ldots m)} + e_{iu}, \tag{1}$$

where μ is a constant and $a_{i_1}^{(1)}, \ldots, a_{i_m}^{(m)}, a_{i_1 i_2}^{(12)}, \ldots, a_{i_{m-1} i_m}^{(m-1,m)}, \ldots, a_{i_1 i_2 \ldots i_m}^{(12\ldots m)}, e_{iu}$ are independently normally distributed with zero means and variances

$$\alpha_{10\ldots00}, \ldots, \alpha_{00\ldots01}, \alpha_{11\ldots00}, \ldots, \alpha_{00\ldots11}, \ldots, \alpha_{11\ldots11}, \alpha_e$$

respectively. Thus if T be the set of m-component nonnull (0, 1)-vectors, then the variance components are α_x, for $x \in T$, and also α_e.

In most practical situations, the total number of observations

$$\sum n_i = n, \tag{2}$$

is fixed. The method of unweighted squares of means will be used for the estimation of variance components. This is an easily calculated analysis that can be used when all the sub-most cells are filled, and the means of these cells are used as observations and

subjected to a balanced data analysis, as suggested by Yates (1934); see Searle (1971, p. 365). In order that mean squares for all main effects, interactions and residual can be calculated

$$s_j \geq 2 \quad (1 \leq j \leq m), \tag{3a}$$

all the $\Pi\, s_j$ cells must have at least one observation, and at least one cell must have more than one observations, so that

$$n > \prod_{j=1}^{m} s_j. \tag{3b}$$

With other methods of estimation, not all the cells necessarily contain observations.

This paper considers in two stages the problem of selection of decision variables $s_1, \ldots, s_m$ and $\{n_i\}$, subject to the constraints (2), (3a) and (3b), for the optimal estimation of variance components. In the first stage, the optimal choice of the cell frequencies $\{n_i\}$, for fixed $\{s_j\}$, has been analytically investigated. The second stage, which is mostly prospective, allows the $\{s_j\}$ also to vary and indicates an approach concerning derivation of minimax designs.

2. Optimal selection of the cell frequencies

For each i, let $\bar{Y}_i = n_i^{-1} \Sigma\, Y_{iu}$, where the summation is over u, be the mean of the ith cell. Let $\bar{Y}$ be a $v \times 1$ vector, where $v = \Pi\, s_j$, with elements $\bar{Y}_i$'s, lexicographically ordered. Denote by D a $v \times v$ diagonal matrix with diagonal elements given by the reciprocals of the cell frequencies $\{n_i\}$ and let $c = v^{-1} \operatorname{tr}(D)$. For $1 \leq j \leq m$, let I_j be the $s_j \times s_j$ identity matrix, E_j the $s_j \times s_j$ matrix with all elements unity,

$$V_j^{x_j} = \begin{cases} I_j & (x_j = 1), \\ E_j & (x_j = 0); \end{cases} \quad W_j^{x_j} = \begin{cases} I_j - s_j^{-1} E_j & (x_j = 1), \\ s_j^{-1} E_j & (x_j = 0). \end{cases} \tag{4a}$$

For $x = (x_1, \ldots, x_m) \in T$, define

$$\tau(x) = \prod_{j=1}^{m} s_j^{x_j}, \quad \beta(x) = \prod_{j=1}^{m} (s_j - 1)^{x_j}, \quad V^x = \bigotimes_{j=1}^{m} V_j^{x_j}, \quad W^x = \bigotimes_{j=1}^{m} W_j^{x_j}, \tag{4b}$$

where $\otimes$ denotes a Kronecker product.

For each $x = (x_1, \ldots, x_m) \in T$, the sum of squares corresponding to the factorial effect $F_1^{x_1} \ldots F_m^{x_m}$ is given by $S_x = \bar{Y}' W^x \bar{Y}$, while that corresponding to error is $S_e = \Sigma\Sigma\, (Y_{iu} - \bar{Y}_i)^2$. Define the mean sum of squares $M_x = S_x / \beta(x)$, for $x \in T$, and $M_e = S_e / (n - v)$. Under the model (1), $\bar{Y}$ is multivariate normal with covariance matrix

$$V = \sum_{x \in T} \alpha_x V^x + \alpha_e D. \tag{5}$$

Hence if one defines $T_x = \{y = (y_1, \ldots, y_m)\colon y \in T,\ y_j \geq x_j,\ 1 \leq j \leq m\}$, and observes that by (4a) and (4b), for any $y, q \in T$,

$$W^y V^q = \begin{cases} v\{\tau(q)\}^{-1} W^y & (q \in T_y), \\ 0 & (q \notin T_y), \end{cases} \tag{6}$$

then it follows that

$$E(M_x) = v \sum_{y \in T_x} \{\tau(y)\}^{-1} \alpha_y + c\alpha_e \quad (x \in T), \quad E(M_e) = \alpha_e. \tag{7}$$

Let x^* be the member of T having all elements unity. Then by (7), unbiased estimators of the variance components are obtained as

$$\hat{\alpha}_x = v^{-1}\tau(x)(-1)^{\Sigma x_i} \sum_{y \in T_1} (-1)^{\Sigma y_i} M_y, \quad x \in T \quad (x \neq x^*),$$
$$\hat{\alpha}_{x^*} = M_{x^*} - cM_e, \quad \hat{\alpha}_e = M_e. \tag{8}$$

By standard results on the covariances of quadratic forms in a multivariate normal setting (Graybill, 1961),

$$\operatorname{cov}(M_y, M_z) = 2\{\beta(y)\beta(z)\}^{-1} \operatorname{tr}(W^y V W^z V), \tag{9}$$

for $y, z \in T$, where V is as in (5). Now, by (5) and (6), $W^y V = f_\alpha(y) W^y + \alpha_e W^y D$, where

$$f_\alpha(y) = v \sum_{q \in T_1} \{\tau(q)\}^{-1} \alpha_q. \tag{10}$$

Also, by (4a) and (4b), $\operatorname{tr}(W^y) = \beta(y)$, $\operatorname{tr}(W^y D) = v^{-1}\beta(y) \operatorname{tr}(D) = c\beta(y)$ and $W^y W^z = 0$ for $y \neq z$. Hence by (9), for $y, z \in T$ $(y \neq z)$,

$$\operatorname{var}(M_y) = 2\{\beta(y)\}^{-1}[\{f_\alpha(y)\}^2 + 2cf_\alpha(y)\alpha_e + \alpha_e^2\{\beta(y)\}^{-1}h(y, y)],$$
$$\operatorname{cov}(M_y, M_z) = 2\{\beta(y)\beta(z)\}^{-1}\alpha_e^2 h(y, z),$$

where $h(y, z) = \operatorname{tr}(W^y D W^z D)$. Also, trivially, $\operatorname{var}(M_e) = 2\alpha_e^2/(n - v)$, $\operatorname{cov}(M_y, M_e) = 0$ $(y \in T)$. Hence by (8), after some simplification,

$$\operatorname{var}(\hat{\alpha}_x) = g_\alpha(x), \quad x \in T \quad (x \neq x^*),$$
$$\operatorname{var}(\hat{\alpha}_{x^*}) = g_\alpha(x^*) + 2c^2\alpha_e^2/(n - v), \quad \operatorname{var}(\hat{\alpha}_e) = 2\alpha_e^2/(n - v), \tag{11}$$

where, for $x \in T$,

$$g_\alpha(x) = 2\{v^{-1}\tau(x)\}^2 \left[\sum_{y \in T_1} \frac{f_\alpha(y)}{\beta(y)} \{f_\alpha(y) + 2c\alpha_e\} + \alpha_e^2 \sum\sum_{y, z \in T_x} \frac{(-1)^{\Sigma(y_i + z_i)}}{\beta(y)\beta(z)} h(y, z) \right]. \tag{12}$$

For fixed $s_1, \ldots, s_m$, under (2), the arithmetic mean-harmonic mean inequality shows that c is minimum when the $\{n_i\}$ are all equal. Also, by the lemma in the Appendix, for each $x \in T$,

$$\sum\sum_{y, z \in T_x} (-1)^{\Sigma(y_i + z_i)} \{\beta(y)\beta(z)\}^{-1} h(y, z)$$

is a minimum when the $\{n_i\}$ are equal. Hence we have the following.

THEOREM 1. *For fixed $s_1, \ldots, s_m$, under (2), for each $x \in T$, $\operatorname{var}(\hat{\alpha}_x)$ is minimized, uniformly in the unknown parameters α_x, $x \in T$, and α_e, when the $\{n_i\}$ are all equal.*

3. THE MINIMAX DESIGN

This section considers the optimal selection of the $\{s_j\}$. In view of Theorem 1, it is reasonable to compare the different choices of the $\{s_j\}$ taking the $\{n_i\}$ all equal. Then for any fixed $\{s_j\}$, $D = vn^{-1}I$, where I is the $v \times v$ identity matrix, $c = vn^{-1}$, $h(y, y) = v^2 n^{-2}\beta(y)$, $h(y, z) = 0$ $(y \neq z)$, and by (12), for $x \in T$,

$$g_\alpha(x) = 2\{v^{-1}\tau(x)\}^2 \sum_{y \in T_x} \{\beta(y)\}^{-1}\{f_\alpha(y) + vn^{-1}\alpha_e\}^2. \tag{13}$$

The variances of $\hat{\alpha}_x$ $(x \in T)$ and $\hat{\alpha}_e$ are now given by (11) and (13).

Two major problems arise in the optimal selection of the $\{s_j\}$. First, for any particular $x \in T$, there is no single choice of the $\{s_j\}$ that minimizes var $(\hat{\alpha}_x)$ uniformly in the unknown parameters. Secondly, for any fixed set of values of the unknown parameters, no single choice of the $\{s_j\}$ minimizes each of var $(\hat{\alpha}_x)$ $(x \in T)$ and var $(\hat{\alpha}_\epsilon)$.

To overcome the second difficulty, one may minimize

$$\rho = \sum_{x \in T} w_x \text{ var } (\hat{\alpha}_x) + w_\epsilon \text{ var } (\hat{\alpha}_\epsilon), \tag{14}$$

where w_x $(x \in T)$ and w_ϵ are known nonnegative weights. In particular, under the A-optimality criterion, these weights are all equal. Difficulty, however, remains as no single choice of the $\{s_j\}$ can minimize ρ uniformly in the unknown parameters. Hence, bringing in the minimax criterion, suppose, as happens in most practical situations,

$$\alpha_x \leq \xi_x \quad (x \in T), \quad \alpha_\epsilon \leq \xi_\epsilon, \tag{15}$$

where ξ_x $(x \in T)$ and ξ_ϵ are known positive quantities. Clearly, the maximum of ρ, subject to (15), becomes ρ_ξ obtained taking $\alpha_x = \xi_x$ $(x \in T)$ and $\alpha_\epsilon = \xi_\epsilon$.

One may now attempt to select the $\{s_j\}$, subject to (3a) and (3b), so as to minimize ρ_ξ. The resulting minimax strategy will also be locally optimal for $\alpha_x = \xi_x$ $(x \in T)$ and $\alpha_\epsilon = \xi_\epsilon$; for a similar situation in the context of minimum normed quadratic unbiased estimation theory, see Rao (1973, Ch. 4). It is, indeed, hard to obtain the solution analytically, since the $\{s_j\}$ have to be integer valued and, as a function of the $\{s_j\}$, ρ_ξ is involved. At least for moderate values of the total sample size n, numerical methods, essentially based on a complete enumeration, are successful. Let ρ_ξ attain a minimum when the $\{s_j\}$ equal $\{s_{j0}\}$. If n is an integral multiple of $\Pi\, s_{j0}$ then a balanced design exists for the choice $\{s_{j0}\}$, and this will be minimax over all possible selections of the $\{s_j\}$ and $\{n_i\}$. If n is not an integral multiple of $\Pi\, s_{j0}$ then such a balanced design does not exist but a nearly balanced design corresponding to $\{s_{j0}\}$ can be expected to be highly satisfactory. As a referee remarks, in the latter situation a subset of a balanced design may be optimal. Incidentally, the optimal designs of Mostafa (1967) were nearly balanced.

Example 1. For $m = 2$, it follows from (10), (11), (13) and (14) that

$$\begin{aligned}\rho_\xi = {} & 2w_{01}\{s_1^2(s_2-1)\}^{-1}(s_1\xi_{01}+\xi_{11}+vn^{-1}\xi_\epsilon)^2+2w_{10}\{s_2^2(s_1-1)\}^{-1}(s_2\xi_{10}+\xi_{11}+vn^{-1}\xi_\epsilon)^2 \\ & +2\{(s_1-1)(s_2-1)\}^{-1}(w_{01}s_1^{-2}+w_{10}s_2^{-2}+w_{11})(\xi_{11}+vn^{-1}\xi_\epsilon)^2 \\ & +2(n-v)^{-1}(w_{11}v^2n^{-2}+w_\epsilon)\xi_\epsilon^2.\end{aligned}$$

With $n = 50$, $w_{01} = w_{10} = 0{\cdot}35$, $w_{11} = 0{\cdot}20$, $w_\epsilon = 0{\cdot}10$, $\xi_{10} = \xi_{01} = 3$, $\xi_{11} = 2$, $\xi_\epsilon = 1$, it may be shown that ρ_ξ is minimum, subject to (3a) and (3b), when $s_1 = 6$, $s_2 = 8$, the corresponding value of ρ being say, $\rho^* = 3{\cdot}20236$. Clearly, for no choice of the $\{s_j\}$ and $\{n_i\}$, subject to (2), (3a) and (3b), the maximum of ρ, over the range (15), can be less than ρ^*. For $n = 50$, $s_1 = 6$, $s_2 = 8$, a balanced design does not exist but nearly balanced designs are found to be highly satisfactory. For example, for the design d with $s_1 = 6$, $s_2 = 8$, $n_{00} = n_{11} = 2$, $n_{i_1 i_2} = 1$ for every other (i_1, i_2), by (10), (11), (12) and (14), the maximum of ρ, under (15), becomes 3·21577. Comparison with ρ^* shows that the efficiency of d is as high as 99·58%.

4. CONCLUDING REMARKS

It would be interesting to find optimal designs, under a general m-factor setting, using other methods of estimation; for details see Anderson (1975, 1981), Thompson &

Anderson (1975) and Muse & Anderson (1978) who present, among other things, interesting results on optimal designs for nested random effects models. It will also be interesting to know about a best combination of estimator and design in general settings.

Another important practical problem is to find the optimal design when the cost per observation varies from cell to cell and (2) is replaced by a restriction on total cost. Then, even with the method of unweighted squares of means, an optimal choice of the cell frequencies, uniformly in the parameters in the sense of Theorem 1, is not possible. Therefore analytical solutions will be hard to obtain, numerical studies may yield interesting results.

Acknowledgement

We are grateful to Professor D. R. Cox for posing the problem and to him and the referees for highly constructive suggestions.

Appendix

A lemma and its proof

Lemma. *For fixed* $\{s_j\}$, *under* (2), $\Sigma\Sigma\, a_y a_z h(y, z)$ *is a minimum when the* $\{n_i\}$ *are all equal provided that the* a_y, a_z $(y, z \in T)$ *are free from the* $\{n_i\}$.

Proof. From the definitions of D and $h(y, z)$ $(y, z \in T)$, one obtains

$$\sum_{y,z\in T}\sum a_y a_z h(y, z) = \operatorname{tr}(WDWD) = \sum\sum w_{ij}^2 (n_i n_j)^{-1},$$

where $W = ((w_{ij})) = \Sigma\, a_y W^y$, for $i = (i_1, \ldots, i_m)$, $j = (j_1, \ldots, j_m)$. Now, W is symmetric and, by (4a) and (4b), $WW = \Sigma\, a_y^2 W^y$. Hence making use of (2),

$$\textstyle\sum_j w_{ij}^2 = w_0, \quad \sum_i w_{ij}^2 = w_0, \quad \sum\sum w_{ij}^2 = v w_0, \quad \sum\sum w_{ij}^2 (n_i + n_j) = 2n w_0,$$

where $w_0 = \Sigma\, a_y^2 \beta(y)/v$. By Hölder's inequality (Rao, 1973, p. 55),

$$\sum\sum w_{ij}^2 (n_i n_j)^{-1} \geq 4 \sum\sum w_{ij}^2 (n_i + n_j)^{-2} \geq 4 (\sum\sum w_{ij}^2)^3 \{\sum\sum w_{ij}^2 (n_i + n_j)\}^{-2} = v^3 n^{-2} w_0,$$

under (2), with equality when the $\{n_i\}$ are all equal.

References

Ahrens, H. & Pincus, R. (1981). On two measures of unbalancedness in a one-way model and their relation to efficiency. *Biometrical J.* **23**, 227-35.

Anderson, R. L. (1975). Designs and estimators for variance components. In *A Survey of Statistical Design and Linear Models*, Ed. J. N. Srivastava, pp. 1-29. Amsterdam: North-Holland.

Anderson, R. L. (1981). Recent developments in designs and estimators for variance components. In *Statistics and Related Topics*, Ed. M. Csörgő, D. A. Dawson, J. N. K. Rao and A. K. Md. E. Saleh, pp. 3-22. Amsterdam: North-Holland.

Graybill, F. A. (1961). *An Introduction to Linear Statistical Models.* New York: McGraw-Hill.

Khuri, A. I. & Sahai, H. (1985). Variance components analysis: a selective literature survey. *Int. Statist. Rev.* **53**, 279-300.

Kurkjian, B. & Zelen, M. (1963). Applications of the calculus for factorial arrangements I. Block and direct product designs. *Biometrika* **50**, 63-73.

Mostafa, M. G. (1967). Designs for simultaneous estimation of functions of variance components from two-way crossed classifications. *Biometrika* **54**, 127-31.

Muse, H. D. & Anderson, R. L. (1978). Comparison of designs to estimate variance components in a two-way classification model. *Technometrics* **20**, 159-66.

PATTERSON, H. D. & THOMPSON, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**, 545–54.

RAO, C. R. (1973). *Linear Statistical Inference and its Applications*, 2nd ed. New York: Wiley.

SEARLE, S. R. (1971). *Linear Models*. New York: Wiley.

THOMPSON, W. O. & ANDERSON, R. L. (1975). A comparison of designs and estimators for the two-stage nested random model. *Technometrics* **17**, 37–44.

YATES, F. (1934). The analysis of multiple classifications with unequal numbers in the different classes. *J. Am. Statist. Assoc.* **29**, 51–66.

[*Received October* 1985. *Revised August* 1987]