A CLASS OF FACTORIAL DESIGNS WITH UNEQUAL CELL-FREQUENCIES

By Gideon Schwarz[1]

Columbia University

1. Summary. A class of multifactorial designs are defined and analyzed. The designs considered have each a total number of observations that can not be divided equally among the cells of the designs; however, by distributing the observations in a way that is in a certain sense symmetrical, the equations that determine the least squares estimates of the linear parameters become explicitly solvable.

The case of two non-interacting factors with arbitrary numbers of levels is treated first. In the n-factor case we have to restrict ourselves to factors having equal numbers of levels. After defining the designs, the estimates are computed. Some general discussions of the symmetries and algebraic properties involved conclude the paper.

2. Introduction. The first case to be considered is that of two non-interacting factors, with I and J levels respectively. For each pair i, j of levels the measured magnitude has an expected value η_{ij}. We assume that the η_{ij} can be expressed in terms of $I + J + 1$ parameters $\{\mu, \alpha_i, \beta_j\}$ by the equations

$$\eta_{ij} = \mu + \alpha_i + \beta_j, \qquad \alpha. = \beta. = 0. \tag{1}$$

The dot indicates as usual summation over the range of the index it replaces.

Denoting by y_{ijk} the kth measurement in the cell in which the factors A and B are applied at levels i and j respectively, we assume the y_{ijk} to be normal independent random variables with means η_{ij} and common variance σ^2.

The experimenter is free to choose the number n_{ij} of observations in each cell. The choice of the matrix n_{ij} may be influenced by three requirements; first, the cost of experimentation makes an unnecessarily large number of observations undesirable; second, for a given number $n_{..}$ of observations, different ways of dividing this number among the different cells will result in different patterns of information about the parameters, and unless specific conditions about some of the levels are added, the design will be the closer to optimal the more evenly the number $n_{..}$ is distributed among the cells; and last, it is impossible to write simple explicit formulas for the least-squares-estimates that hold for general n_{ij}, while for some classes of n_{ij}-matrices, such formulae can be found.

Considering the two last requirements only, we are led to a well known class of designs, namely those in which all the n_{ij} are equal, say to n.

Received July 17, 1959; revised April 19, 1960.

[1] This research was supported by the Office of Naval Research under Contract Number Nonr-266(59), Project Number 042-205. Reproduction in whole or in part is permitted for any purpose of the United States Government.

As we have $n_{\cdot\cdot} = nIJ$ for this class, we cannot regulate the total number of experiments except in jumps of IJ; in many cases this may lead to a violation of the first requirement. Consider for example a case in which one observation per cell would suffice for estimation of the parameters μ, α_i, β_j, while for the estimation of σ^2, we would want a few additional observations in some of the cells. Within the class of constant n_{ij}, this can be achieved only by doubling the total number of experiments.

There have been various attempts of considering special designs with unequal frequencies (Cf. References). Among the special cases treated by Daniel [2] and, in private communications with Daniel, by A. Birnbaum and Scheffé, there were designs with some symmetry properties. It was Birnbaum's suggestion to look for a more general class of designs that led to the results described in this paper.

3. Definition of S and Calculation of the Estimates. Let us proceed now to define the class S. We start out with d by d unit matrix, d being any common divisor of I and J, and change it into an I by J matrix by replacing each of its "one" entries by a I/d by J/d matrix of ones, and each of its zeros by a similar matrix of zeros. This way we define a matrix

$$\left(\begin{array}{ccc|ccc} 1 & 1 & \cdots & & & \\ 1 & \cdot & & & & \\ 1 & & \cdot & & 0 & \cdots \\ \vdots & & & & & \\ \hline & & & 1 & 1 & \cdots \\ & & & 1 & & \\ & & 0 & 1 & & \\ & & & \vdots & & \\ \cline{4-6} & & \vdots & 0 & & \ddots \end{array}\right) .$$

Denoting this matrix by $A_{I,J,d}$, or for short $A_{I,J}$, we can now define S as the class of all designs with matrices (n_{ij}) that can be written either in the form $(n) + A_{I,J}$ or $(n) - A_{I,J}$ for some positive integer n, and d, a divisor of n, where (n) denotes the I by J matrix having all entries equal to n. We claim, (a) the number $n_{\cdot\cdot}$ runs in the class S over all integers of the form

$$IJ(n \pm d^{-1});$$

and (b) there is a simple explicit formula for the least-squares estimates that holds for all the designs in S.

(a) becomes evident if we observe that $A_{I,J}$ has IJ/d non-zero entries, and we shall prove (b) by arriving at the formulae, first for the minus sign and then in general.

The least squares estimates of the row effects can be obtained from the numbers

$$a_i = y_{i\cdot\cdot}/n_{i\cdot} - y_{\cdot\cdot\cdot}/n_{\cdot\cdot}, \tag{2}$$

which span uniquely the estimation space restricted by the side conditions. Each a_i is a unique linear combination of the least squares estimates $\hat{\alpha}_i$, $\hat{\beta}_i$ given by

$$(3) \qquad a_i = \hat{\alpha}_i + [1/J(n + d^{-1})] \textstyle\sum^* \hat{\beta}_j ,$$

where $\sum^*$ denotes summation over the cells with $n + 1$ observations only. Using vector notation $a = (a_1, \cdots, a_I)$, etc.,

$$(4) \qquad a = \hat{\alpha} + [d/J(nd + 1)]\, A_{IJ}\hat{\beta},$$

and, by interchanging rows and columns,

$$(5) \qquad b = \hat{\beta} + [d/I(nd + 1)]\, A_{JI}\hat{\alpha}.$$

To eliminate $\hat{\beta}$ from equations (4) and (5), we subtract from (4) a suitable multiple of (5). Using the equation

$$(6) \qquad A_{IJ}A_{JK} = (J/d)A_{IK}$$

which follows easily from the definition of A_{IJ}, we arrive at

$$(7) \qquad a - [d/J(nd + 1)]A_{IJ}b = \hat{\alpha} - [d/I(nd + 1)^2]A_{II}\hat{\alpha}.$$

In order to solve this equation, we have to invert a matrix which can be written, if we denote the unit matrix by U_{II}, as $U_{II} - [d/I(nd + 1)^2]A_{II}$.

We can find the required inverse by finding the value of t that makes the product

$$(8) \qquad (U_{II} - [d/I(nd + 1)^2]A_{II})\,(U_{II} + tA_{II})$$

equal to the unit matrix. Reducing the A_{II}^2 term by applying (6) we obtain $t = 1/In(nd + 2)$. Having found the inverse we can now solve equation (7). Denoting by R and C vectors of row and column-sums, respectively, and by S a vector with I components, all equal to the grand total $y_{\cdots}$, we have

$$(9) \qquad \begin{aligned}\hat{\alpha} = [d/J(nd + 1)]R + [d/IJn(nd + 1)(nd + 2)]A_{II}R \\ - [d/IJn(nd + 2)]A_{IJ}C - [d/IJ(nd + 2)]S.\end{aligned}$$

The corresponding formula for β is easily obtained by interchanging R and C, as well as I and J. The estimate of μ is obviously equal to $(d/IJ(nd - 1))y\cdots$, the mean of all observations. The change in the formulae for the case $(n_{ij}) = (n) - A_{I,J}$, will consist of changing the signs of d, and of all the matrices. Merging both cases into one, and denoting by S also a vector with J components all equal to $y\cdots$, we have finally

$$(10) \qquad \begin{aligned}\hat{\alpha} = [d/J(nd \pm 1)]R + [d/IJn(nd \pm 1)(nd \pm 2)A_{II}R \\ \mp [d/IJn(nd \pm 2)]A_{IJ}C - [d/IJ(nd \pm 2)]S,\end{aligned}$$

$$(11) \qquad \begin{aligned}\hat{\beta} = [d/I(nd \pm 1)]C + [d/JIn(nd \pm 1)(nd \pm 2)]A_{JJ}C \\ \mp [d/JIn(nd \pm 2)]A_{JI}R - [d/JI(nd \pm 2)]S,\end{aligned}$$

$$(12) \qquad \hat{\mu} = [d/IJ(nd \pm 1)]S.$$

As a final remark, let us note that our definitions and formulae are valid as long as n is at least 1, and d is at least 2, with the exception of the case $n = 1$, $d = 2$, in which $nd - 2$ equals zero, and the $n - d^{-1}$ replicate is not sufficient for estimation of the parameters. On the other hand, as for $d \geqq 3$, $n = 1$ the formulae remain meaningful also for the lower sign, certain designs with some empty cells are included in the class considered here.

Most of what has been done in the preceeding section admits a rather straightforward generalization to the case of q factors acting additively, that is, with no interactions of any order. The only step that is not generalized so easily is the reduction of equations (4) and (5), each involving both row-effect estimates and column effect estimates, to equation (7), which isolates the row effects. In order to make possible an explicit solution to the analogous problem in the case of many factors, we have to restrict our considerations to designs having an equal number of levels for every factor. Denoting the effect of the hth factor at its ith level by $\alpha_{i(h)}$, we define our model by the equations

$$y_{i,k} = \mu + \sum_h \alpha_{i(h)} + \epsilon_{i,k} \tag{13}$$

where i denotes the vector $(i(1), \cdots i(q))$, and $k = 1, 2, \cdots, n_{i,j}$, with the error terms distributed as usual. About the parameters we assume $\sum_i \alpha_{i(h)} = 0$, $h = 1, 2, \cdots q$.

In order to determine the number of observations in each cell, we choose a divisor d of the number of levels I, and construct a q-dimensional hyper-cube of side-length d. Putting d ones at the grid points along the q-space-diagonal of the hyper-cube, and zeros at the other grid points, we obtain the q-dimensional analogue of the d by d unit matrix. Replacing each $(q - 1)$ dimensional layer by I/d identical layers, an array of I^q points is obtained, I^q/d^{q-1} of which carry units. If we start out with an I^q-design having n observations in each cell, and add ± 1 observation to each cell that corresponds to a unit in the array, an $n \pm 1/d^{q-1}$ duplicate will be obtained.

Defining the numbers $a_i(h)$ as the average of the observations in the layer determined by a given level of a given factor, minus the average of all observations, we get a system of vector equations;

$$\begin{aligned} a(1) &= \hat{\alpha}(1) + gA_{II}\hat{\alpha}(2) + gA_{II}\hat{\alpha}(3) + \cdots + gA_{II}\hat{\alpha}(q) \\ a(2) &= gA_{II}\hat{\alpha}(1) + \hat{\alpha}(2) + gA_{II}\hat{\alpha}(3) + \cdots + gA_{II}\hat{\alpha}(q) \\ &\vdots \qquad\qquad\qquad\qquad\qquad \vdots \\ a(q) &= gA_{II}\hat{\alpha}(1) + \cdots \qquad \cdots + gA_{II}\hat{\alpha}(q-1) + \hat{\alpha}(q) \end{aligned} \tag{14}$$

where $g = [d^{q-1}/I^{q-1}(nd^{q-1} \pm 1)](I/d)^{q-2} = d/I(nd^{q-1} \pm 1)$. The first factor in g is the reciprocal of the number of observations per $(n - 1)$ − dimensional partial design. The second factor is the number of higher populated cells that two levels of different factors have in common.

For the solution of (14) inversion of a q by q matrix having I by I matrices as elements (a q by q by I by I tensor of the fourth degree) is required. We start by considering the matrix G_{qq} obtained by replacing A_{II} in the tensor of (14) by a scalar variable x. Putting for its inverse

$$G_{qq}^{-1} = \begin{pmatrix} z & y & y & \cdots & y \\ y & z & y & \cdots & y \\ \vdots & \vdots & \vdots & & \vdots \\ y & y & \cdots & y & z \end{pmatrix}$$

we find

$$y = gx/[(q-1)g^2x^2 - (q-2)gx - 1],$$
$$z = -(q-2)gx - 1/[(q-1)g^2x^2 - (q-2)gx - 1].$$

We can write this result in a form that does not involve any fractions

$$G_{qq}[((q-1)g^2x^2 - (q-2)gx - 1)G_{qq}^{-1}] = ((q-1)g^2x^2 - (q-2)gx - 1)U_{qq}, \tag{15}$$

where the expression in the square brackets equals a q by q matrix with $-(q-2)gx - 1$ along the diagonal, and gx in the other places. Having disposed of fractions, we can now substitute A_{II} for x. Carrying out the substitution in equation (15), G_{qq} becomes the tensor of (14), and the expression in the square brackets becomes a tensor having $-(q-2)gA_{II} - U_{II}$ along its diagonal, and gA_{II} in the other places.

Applying all this to (14), we arrive at the reduced equations,

$$[(q-1)g^2A_{II}^2 - (q-2)gA_{II} - U_{II}]\,\hat{\alpha}(1) = -[(q-2)gA_{II} - U_{II}]a(1) + gA_{II}[a(2) + \cdots] \tag{16}$$

We can now proceed as in the two-factor case, and get, putting N for nd^{q-1},

$$\begin{aligned} \hat{\alpha}(1) &= d^{q-1}/[I^{q-1}(N \pm 1)]S(1) \\ &\quad + [(q-1)d^q/I^qN(N \pm 1)(N \pm q)]A_{II}S(1) \\ &\quad \mp d^q/[I^qN(N-q)A_{II}[S(2) + \cdots S(q)] - d^{q-1}/I^q(N \pm q)S. \end{aligned} \tag{17}$$

As in the 2-factor case, the lower signs serve for the $n - 1/d^{q-1}$ duplicate. As for $q \geqq 3$ we have $N = nd^{q-1} > q$, the denominators never vanish except in the case mentioned before when $q = 2$ and $d = 2$, and the formulae are valid unrestrictedly. By permuting the factors, estimates for the other factors can be easily obtained.

4. General Symmetrical Designs. In this section we shall examine closer the symmetry properties that the designs treated in this paper have in common.

The various symmetry properties of the designs having equal numbers of observations in all cells are implied by the invariance of these designs under all permutations of the levels of any factors; furthermore, those designs, the "full multiple replicates", are the only ones left invariant by all permutations. Certainly, invariance under all permutations assures us of equal treatment of all levels of each factor. It implies an even stronger property: different ordered pairs of levels will enter the design similarly, as will any different ordered n-tuples. This additional property is certainly welcome. Some of the questions the designed experiment might be called upon to answer do involve pairs or other sets of levels, and it would be natural to expect symmetric treatment of these questions as well. However, we know that we have to give up some requirements if we want to include fractional replicates, and it is this "symmetry of subsets" that we choose to sacrifice.

Let us examine the freedom gained by requiring symmetry with regard to single levels only, by looking at the two-factor case. In this case, the design is determined by a matrix having the cell frequencies as its entries. Applying single-level symmetry to the row factor, we find that the rows of the matrix have to be equal to each other; however, as the order of entries in a row is determined by the order of levels of the column factor, the order in a row is immaterial, and the word "equal" should be read "differing only by a permutation of their elements". Similar "equality" is implied for the columns of the matrix. Any unit matrix can now serve as an example of a matrix having the required properties and yet not belonging to the full replicate designs.

Returning to the multifactorial designs, we arrive at the following formulation of our symmetry requirements:

Definition: A design is called "symmetrical with regard to single levels", or from here on, for short, "symmetrical", if the two partial designs resulting from fixing any one of the factors at two different levels, can be transformed one into the other by permuting the levels of the other factors.

We now restrict our class of designs even further, by introducing a restriction that is not motivated solely by considerations of symmetry. The designs we shall consider will all have only two different numbers of observations per cell occurring in their cells, furthermore, those two numbers will differ from each other by one. We justify this restriction by the following "optimality argument": the definition of a symmetrical design implies that the different cell frequencies appearing in partial designs belonging to different levels of the same factor, will be the same, possibly differently arranged. If there were two cells in the design whose numbers of observations differed by more than one, we would find two such cells in every subdesign and a new design could be defined by decreasing by one the number of observations in the higher populated cell and increasing it in the other. The resulting design would still be symmetrical and have the same total number of observations as the original design. As whenever the given total number of observations makes equally populated cells possible, the fully replicated design is in some sense optimal, we can interpret the above restriction as an attempt to avoid unnecessary deviations from optimality.

Having narrowed down the class of designs, we can now turn to the last requirement: existence of explicit estimation formulae.

The fact that permitted us to look for an inverse of a linear polynomial in

A_{II} among the set of linear polynomials in A_{II} is the degree of minimal polynomial of A_{II} : it is quadratic. In general, the inverse of any regular matrix P that is a polynomial $P(A)$, where A has a minimal polynomial of degree r, can be written as $Q(A)$, Q being of degree $r - 1$ at most.

PROOF: The set of all such $Q(A)$ is a ring and in this ring the ideal generated by $P(A)$ must be the whole ring, otherwise it would be of lower dimension and $P(A)$ would be singular. Therefore, $P(A)$ has an inverse among the $Q(A)$.

In general for a I by I matrix r can be any number from 1 to I. As we have to find r constants in order to invert the matrix, the inversion can be done simply only for low r. The class **S** can be characterized as the class of matrices having the symmetries and optimality properties mentioned above, and a minimal polynomial of degree $r = 2$.

REFERENCES

[1] J. D. BANKIER AND R. E. WALPOLE, "Components of variance analysis for proportional frequencies", *Ann. Math. Stat.*, Vol. 28 (1957), pp. 742-53.

[2] C. DANIEL, "Fractional replication of industrial experiments", *Transactions of the 1957 National Convention of the American Society for Quality Control*, pp. 229-233.

[3] J. D. FINNEY, "Main effects and interactions", *J. Amer. Stat. Assoc.*, Vol. 43 (1948), pp. 566-71.

[4] K. KISHEN, "Symmetrical unequal block designs", *Sankhyā*, Vol. 5 (1941), pp. 328-348.

[5] K. R. NAIR, "A note on the method of fitting constants for analysis of nonorthogonal data arranged in a double classification", *Sankhyā*, Vol. 5 (1941), pp. 317-28.

[6] H. SCHEFFÉ, *Analysis of Variance*, John Wiley and Sons, New York, 1959.

[7] G. W. SNEDECOR, *Statistical Methods*, 4th Edition, Iowa State College Press, Ames, Iowa, 1946.

[8] F. YATES, "The analysis of multiple classifications with unequal numbers in the different classes", *J. Amer. Stat. Assoc.*, Vol. 29 (1934), pp. 51-66.