Sankhyā : The Indian Journal of Statistics
1985, Volume 47, Series A, Pt. 3, pp. 350-356.

FRACTION SELECTION PROBLEM IN DISCRETE MULTIVARIATE ANALYSIS

By RAHUL MUKERJEE
Indian Statistical Institute

SUMMARY. This paper considers a multiattribute set-up with some controllable and some non-controllable attributes. It is investigated how one can choose a subset of the form combinations of the controllable attributes such that if independent samples are drawn corresponding to the combinations in the subset and classified according to the form combinations of the non-controllable attributes then the procedure allows estimation of the relevant parameters when prior knowledge is available regarding the absence of some of the interactions. The standard log-linear model has been assumed and the connexion with the traditional theory of fractional factorial plans in design of experiments has been explored.

1. Introduction

The problem of analyzing categorical data under the loglinear model with several attributes has received considerable attention in recent years (for detailed references see Haberman (1974, 1978, 1979) and Bishop, Fienberg and Holland (1975)). If some of the attributes are controllable then one may draw independent samples corresponding to the form combinations of these attributes and classify these samples corresponding to the forms of the remaining (non-controllable) ones. If prior knowledge is available regarding the absence some higher order interactions then in such product multinomial sampling, instead of taking samples corresponding to all form combinations of the controllable attributes one might consider only a subset of these form combinations to estimate the relevant parameters.

As a simple example, with only two attributes F_1 and F_2, each dichotomized with forms 0, 1, if F_1 be controllable it is intuitively clear that a single sample drawn corresponding to the form 0 of F_1 and classified according to the forms of F_2 will enable one to estimate the main effect of F_2, provided there is no interaction between F_1 and F_2. This is just analogous to the simple fact in the context of traditional fractional factorial designs (see Kempthorne (1952), Raghavarao (1971)) that in a 2^2 factorial design with factors F_1 and F_2 starting from the defining equation $I = F_1$ it is possible to estimate main effect F_2 provided interaction F_1F_2 is absent.

AMS (1970) *subject classification :* 62F10.

Key words and phrases : Log-linear model ; Orthogonal array ; Product multinomial sampling.

The above example suggests a strong relationship between the fraction selection problem in categorical data analysis with some controllable attributes and the traditional theory of fractional replication in design of experiments. The objective of the present paper is to investigate this interface rigorously. The necessary and sufficient condition under which a subset of form combinations of the controllable attributes ensure the estimability of the relevant parameters has been developed in section 3. In section 4, it is seen that this condition is precisely equivalent to the corresponding one in a traditional fractional factorial setting.

2. Notations and preliminaries

To formalize the ideas, let there be m attributes $F_1, \ldots, F_m$, the j-th attribute having s_j forms $0, 1, \ldots, s_j-1$ $(1 \leqslant j \leqslant m)$. Let $\pi(i)$ denote the probability that a randomly chosen individual corresponds to the form combination $i = (i_1, \ldots, i_m)$, $0 \leqslant i_j \leqslant s_j-1$, $1 \leqslant j \leqslant m$. Let $v = \prod_{j=1}^{m} s_j$, $\tau(i) = \log \pi(i)$ and $\boldsymbol{\tau}^{(v \times 1)}$ a vector with elements $\tau(i)$'s arranged in the lexicographic order. Then defining Ω as the set of m component vectors with elements 0, 1, for each $x = (x_1, \ldots, x_m) \in \Omega$, it is easy (see Kurkjian and Zelen (1963), Mukerjee (1979)) to find a $\Pi(s_j-1)^{x_j} \times v$ matrix P^x with orthonormal rows such that under the standard log-linear model (cf. Birch (1963), Bishop *et al.* (1975)) $P^x \boldsymbol{\tau}$ represents the interaction $F_1^{x_1} \ldots F_m^{x_m}$ among the attributes.

Among the m attributes suppose that the first p are controllable $(1 \leqslant p < m)$ in the sense that corresponding to each fixed form combination of $F_1, \ldots, F_p$ it is possible to draw a random sample which may be classified according to the forms of $F_{p+1}, \ldots, F_m$. Then for $0 \leqslant i_j \leqslant s_j-1$; $1 \leqslant j \leqslant m$, the conditional probability $\pi(i^{(2)}/i^{(1)})$ that a randomly chosen individual belongs to the form combination $i^{(2)} = (i_{p+1}, \ldots, i_m)$ of $F_{p+1}, \ldots, F_m$ given that it corresponds to the combination $i^{(1)} = (i_1, \ldots, i_p)$ of $F_1, \ldots, F_p$ is given by

$$\pi(i^{(2)}/i^{(1)}) = \frac{\pi(i)}{\sum_{i^{(2)}} \pi(i)} = \frac{e^{\tau(i)}}{\sum_{i^{(2)}} e^{\tau(i)}}. \qquad \ldots \quad (2.1)$$

Defining $\Omega^* = \{(x_1, \ldots, x_m) : x_{p+1} = \ldots = x_m = 0, x_j = 0, 1, 1 \leqslant j \leqslant p\}$ and $\bar{\Omega} = \Omega - \Omega^*$, it may be seen that (2.1) depends on $\boldsymbol{\tau}$ only through $P^x\boldsymbol{\tau}$ for $x \in \bar{\Omega}$. In other words (2.1) is free from the parameters representing the effects $F_1^{x_1} \ldots F_m^{x_m}$ for $(x_1, \ldots, x_m) \in \Omega^*$. Hence under the controlled sampling procedure one can possibly infer only on $P^x\tau$ for $x = (x_1, \ldots, x_m) \in \bar{\Omega}$. Suppose now prior knowledge is available regarding the absence of some of the interactions, say, let it be known that

$$P^x\tau = 0 \text{ for } x \in \Omega_H \ (\subset \bar{\Omega}). \qquad \ldots \quad (2.2)$$

Then defining $B = [\ldots, P^{x\prime}, \ldots]'$ for $x \in \bar{\Omega} - \Omega_H$ and $\Phi = B\tau$, (2.1) reduces to

$$\pi(i^{(2)}/i^{(1)}) = e^{h_i'\Phi} / \left(\sum_{i^{(2)}} e^{h_i'\Phi} \right), \qquad \ldots \quad (2.3)$$

where the h_i's denote the columns of B.

In view of (2.3), from the controlled sampling procedure, under (2.2), one may proceed to infer on Φ i.e. on $P^x\tau$ for $x \in \bar{\Omega} - \Omega_H$. In the next section it will be examined how to choose a subset S of form combinations of $F_1, \ldots, F_p$ so that a controlled sampling based on only the form combinations in S enables one to make such inference.

3. Adequacy of a subset

Consider a subset S of form combinations of $F_1, \ldots, F_p$ and suppose independent samples are drawn only corresponding to these combinations. For $i^{(1)} = (i_1, \ldots, i_p) \in S$, let $n(i^{(1)})$ be the sample size and $n(i^{(2)}/i^{(1)})$ be the observed frequency of the form combinaton $i^{(2)} = (i_{p+1}, \ldots, i_m)$ of $F_{p+1}, \ldots, F_m$ $(0 \leqslant i_j \leqslant s_j - 1,\ p+1 \leqslant j \leqslant m)$. Then by (2.3), under (2.2), the likelihood function in terms of Φ is

$$L = \text{constant} \times \prod_{i^{(1)} \in S} \left[\prod_{i^{(2)}} \left(\frac{e^{h_i'\Phi}}{\sum_{i^{(2)}} e^{h_i'\Phi}} \right)^{n(i^{(2)}/i^{(1)})} \right] \qquad \ldots \quad (3.1)$$

where the constant does not depend on Φ.

Instead of (3.1), however, an equivalent alternative form of the likelihood function will be considered. This will somewhat simplify the presentation.

Let T denote the set of $\prod_{j=p+1}^{m} s_j$ possible choices of $i^{(2)} = (i_{p+1}, \ldots, i_m)$. Then for any $i = (i^{(1)}, i^{(2)})$, where $i^{(1)} \in S$, $i^{(2)} \in T$, let $\mu(i) = \mu(i^{(1)}, i^{(2)}) = h_i'\Phi$. For each $i^{(1)} \in S$, let $H(i^{(1)})$ be a matrix with columns given by h_i's i.e. $h_{i_1 \ldots i_m}$'s arranged lexicographically according to $i^{(2)} = (i_{p+1}, \ldots, i_m)$ and similarly $\mu(i^{(1)})$ be a vector with elments $\mu(i^{(1)}, i^{(2)})$ lexicographically ordered according to $i^{(2)}$. Further, define $H = [\ldots, H(i^{(1)}), \ldots]$ for $i^{(1)} \in S$ and $\mu = [\ldots, \mu(i^{(1)})', \ldots]'$ for $i^{(1)} \in S$. Then

$$\mu = H'\Phi, \qquad \ldots \quad (3.2)$$

i.e. $\mu \in$ column space $(H') = [= \mathscr{M}$, say]. In terms of μ, L as in (3.1) may be written as

$$L = \text{constant} \times \prod_{i^{(1)} \in S} \left[\prod_{i^{(2)}} \left(\frac{e^{\mu(i)}}{\sum_{i^{(2)}} e^{\mu(\bar{i})}} \right)^{n(i^{(2)}/i^{(1)})} \right] \quad \ldots \quad (3.3)$$

where $\mu \in \mathscr{M}$.

Let n be a vector with elements $n(i^{(2)}/i^{(1)})$ formed in the manner in which μ was formed from $\mu(i^{(1)}/i^{(2)})$'s. Then following Theorem 2.2 of Haberman (1974) a necessary and sufficient condition for the existence of a maximum likelihood estimator (MLE) of μ is obtained as :

Theorem 3.1 : *A necessary and sufficient condition that the MLE of* μ *exists is that there exists* $z \in \mathscr{M}^{\perp}$ *such that* $n+z > 0$ *(where* $\mathscr{M}^{\perp}$ *is the orthocomplement of* $\mathscr{M}$ *in the appropriate dimensional Euclidian space).*

Since, by (3.2), our parameters of interest namely Φ are closely linked with μ, the above theorem suggests that the existence of MLE of Φ depends not only on the choice of S but also on the observed cell frequencies (in particular, clearly if the condition stated in Theorem 3.1 does not hold then MLE of Φ cannot exist). Thus the status of any S in this regard should be judged keeping this fact in mind.

Definition 3.1 : A subset S of the form combinations of $F_1, \ldots, F_p$ will be called *adequate* for Ω_H if, given that the condition stated in Theorem 3.1 holds, unique MLE of Φ is available.

The following theorem gives a necessary and sufficient condition for the adequacy of a subset.

Theorem 3.2 : *A subset* S *is adequate for* Ω_H *if and only if the rows of the matrix* H *are linearly independent.*

Proof : In view of Theorem 3.1, the proof is immediate noting that by (3.2) μ is a one-one function of Φ when and only when H has linearly independent rows.

4. Connexion with fractional factorial plans

As indicated in the introduction, the last result has a close link with the theory of fractional factorial plans in design of experiments.

Corresponding to the multiattribute setting described in section 2, define the following. Let $F_1, \ldots, F_m$ be m factors at $s_1, \ldots, s_m$ levels (e.g. the levels are $0, 1, \ldots, s_j - 1$ for F_j), there being $v = \prod_{j=1}^{m} s_j$ level combinations in all.

Denote by $\tau^{(v\times 1)}$ the vector of treatment effects arranged lexicographically. For $x = (x_1, \ldots, x_m) \varepsilon \Omega$, $P^x\tau$ represents a full set of orthonormal contrasts belonging to the factorial effect $F_1^{x_1} \ldots F_m^{x_m}$ (cf. Kurkjian and Zelen (1963) Mukerjee (1979)), where if, in particular, $x_1 = \ldots = x_m = 0$, then $F_1^{x_1} \ldots F_m^{x_m}$ stands for the 'general effect'.

Let S be a subset of level combinations of $F_1, \ldots, F_p$. Then by an *equireplicate fraction generated from* S, we mean a design comprising the level combinations $\{(i_1, \ldots, i_m) : (i_1, \ldots, i_p) \varepsilon S,\ 0 \leqslant i_j \leqslant s_j - 1$ for $p+1 \leqslant j \leqslant m\}$ in which the number of replications of $(i_1, \ldots, i_m)$ depends only on $(i_1, \ldots, i_p)$. Using the notations of sections 2,3, denote the number of replications of the level combination $i = (i^{(1)}, i^{(2)})$ in such a fraction by $\nu(i^{(1)})$ (> 0), $i^{(1)} \varepsilon S$, $i^{(2)} \varepsilon T$.

Define $\bar{\Omega}$, Ω^*, Ω_H as in Section 2. Suppose, analogously to (2.2),

$$P^x\tau = 0 \text{ for } x \epsilon \Omega_H \qquad \ldots \quad (7.1)$$

Defining B and $H(i^{(1)})$, $i^{(1)} \varepsilon S$, as in sections 2 and 3, it may be seen after considerable algebra that the least squares reduced normal equations for $B\tau$ under (4.1), based on an equireplicate fraction generated from S, have a coefficient matrix given by

$$\sum_{i^{(1)} \varepsilon S} \nu(i^{(1)}) H(i^{(1)}) H(i^{(1)})'. \qquad \ldots \quad (4.2)$$

The derivation of (4.2), which follows the line of Chakrabarti (1962, Ch. 2) and Mukerjee (1980), is lengthy and hence omitted. For the interested reader reference is made to Mukerjee (1984) in this regard.

Since $\nu(i^{(1)}) > 0$, for each $i^{(1)} \varepsilon S$, (4.2) is positive definite (so that $B\tau$ is estimable) if and only if the matrix $H = [\ldots, H(i^{(1)}), \ldots]$ for $i^{(1)} \varepsilon S$, as in section 3, has linearly independent rows. This is precisely equivalent to the condition considered in Theorem 3.2. Thus one gets the following theorem :

Theorem 4.1 : *A subset S of the form combinations of $F_1, \ldots, F_p$ in the multiattribute setting is adequate for Ω_H if and only if an equireplicate fraction generated from S ensures the estimability of $B\tau$ in terms of the corresponding factorial set-up.*

This theorem links up the problems of fraction selection in discrete multivariate analysis and in ordinary factorial designs. The existing results on fractional factorial plans may thus be utilized in obtaining adequate subsets for the original problem. In particular, applying Theorem 4.1 and proceeding as in Rao (1947, 1973) we have the following :

Theorem 4.2 : *Let* $\Omega_H = \{(x_1, \ldots, x_m) : (x_1, \ldots, x_m) \varepsilon \bar{\Omega}$, *among* $x_1, \ldots x_p$, *more than u are equal to unity} and S be such that writing the form combinations of S as columns, the resulting array is an orthogonal array of strength* $2u$. *Then S is adequate.*

Example 4.1 : Consider the data on recurrence of rheumatic fever from Bishop *et al.* (1975, pp. 116–119) with five attributes, namely,

F_1 : Laboratory results,

F_2 : Interval from last rheumatic fever attack,

F_3 : Heart disease,

F_4 : Number of previous attacks, and

F_5 : Recurrence of rheumatic fever,

with 4, 2, 2, 2 and 2 forms respectively.

It appears that the first four attributes are controllable and suppose prior information is available regarding the absence of all interactions involving three or more attributes. Then $m = 5$, $p = 4$, $s_1 = 4$, $s_2 = s_3 = s_4 = s_5 = 2$, $\Omega_H = \{(x_1, x_2, x_3, x_4, 1) : x_j = 0, 1;\ j = 1, 2, 3, 4,$ among x_1, x_2, x_3, x_4 more than one equal unity$\}$. Taking

$$S = \{(0, 0, 0, 0), (0, 1, 1, 1), (1, 0, 0, 1)\ (1, 1, 1, 0), (2, 0, 1, 0), (2, 1, 0, 1), (3, 0, 1, 1), (3, 1, 0, 0)\},$$

an application of Theorem 4.2, with $u = 1$, shows S to be adequate for Ω_H. In other words, given that interactions involving three or more attriibutes are absent, a product multinomal sampling corresponding to only eight of the thirty two possible form combinations of F_1, F_2, F_3, F_4 will be adequate for drawing inference on main effect F_5 and interactions $F_jF_5 (j = 1, 2, 3, 4)$.

Acknowledgment

The author wishes to record his deep gratitude to Professor J. K. Ghosh, Indian Statistical Institute, for proposing the problem and for numerous helpful suggestions. Thanks are also due to Dr. A. B. Raha and Dr. Bikas K. Sinha, Indian Statistical Institute, and the referee for their constructive suggestions.

References

Birch, M. M. (1963) : Maximum likelihood in three-way contingency tables. *J. Roy. Statist. Soc.*, Ser. B, **25**, 220 233.

Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1975) : *Discrete Multivariate Analysis: Theory and Practice*, MIT Press, Cambridge, Massachusetts.

Chakrabarti, M. C. (1962) : *Mathematics of Design and Analysis of Experiments*, Asia Publishing, Bombay.

HABERMAN, S. J. (1974): *The Analysis of Frequency Data*, The Univ. of Chicago Press, Chicago.

——— (1978): *Analysis of Qualitative Data*, Vol. 1 Academic Press, New York.

——— (1979): *Analysis of Qualitative Data*, Vol. 2. Academic Press, New York.

KEMPTHORNE, O. (1952): *The Design and Analysis of Experiments*, John Wiley, New York.

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

MUKERJEE, R. (1979): Inter-effect-orthogonality in factorial experiments. *Calcutta Statist. Assoc. Bull.*, 28, 83-108.

——— (1980): Orthogonal fractional factorial plans. *Calcutta Statist. Assoc. Bull.*, 29, 143-160.

——— (1984): Fraction selection problem in discrete multivariate analysis. *Tech. Rep.* 9/84 (unpublished), Indian Statistical Institute, Calcutta.

RAGHAVARAO, D. (1971): *Constructions and Combinatorial Problems in Design of Experiments*, John Wiley, New York.

RAO, C. R. (1947): Factorial experiments derivable from combinatorial arrangements of arrays. *J. Roy. Statist. Soc.*, Ser. B, 9, 128-140.

——— (1973): Some combinatorial problems of arrays and applications to design of experiments. In *A Survey of Combinatorial Theory*, J. N. Srivastava ed., 349-359, North Holland, Amsterdam.

Paper received : May, 1984.

Revised : February, 1985.