J. Ind. Soc. Agril. Statist.
58(2), 2004 : 212-222

A Note on the Nearest Proportional to Size Sampling Design

Raghunath Arnab
University of Durban-Westville, Durban–4000, South Africa
(Received : April, 2003)

SUMMARY

The concept of proportional to size sampling design, nearest to a given sampling design, was introduced by Gabler (1987). Adhikary (1996) provided a set of sufficient conditions for realizations of such a sampling design and a method of construction of rejective π^*ps sampling plan. It is shown in this paper that all the sufficient conditions, laid down by Adhikary (1996), are incorrect. In addition, some new sufficient conditions are introduced.

Key words : Proportional to size sampling design, Poisson sampling scheme, Positive solution, Unique solution.

1. Introduction

Consider a finite population U = {1, ..., I, ..., N} of N identifiable units. Let a sample s of size n be selected from U with probability p(s) according to a sampling design p. The set of all possible samples of size n will be denoted by S. The support T(p), of the sampling design p, is the collection of the sample {s} with $p(s) > 0$ and $\sum_s p(s) = 1$. The class of sampling design p of fixed sample size n will be denoted by P_n.

The inclusion probabilities for the i^{th}, and ij^{th} $(i \neq j)$ units will be denoted by $\pi_i = \pi_{ii} = \sum_{s \supset i} p(s) = \sum_{s \in S} I_{si} p(s)$ and $\pi_{ij} = \sum_{s \supset ij} p(s) = \sum_{s \in S} I_{si} I_{sj} p(s)$ respectively, where $I_{si} = 1(0)$ if $i \in S\ (i \notin S)$. The inclusion probability matrix of the design p is denoted as

$$\pi = \begin{pmatrix} \pi_{11} & \cdot & \pi_{1j} & \cdot & \pi_{1N} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \pi_{i1} & \cdot & \pi_{ij} & \cdot & \pi_{iN} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \pi_{N1} & \cdot & \pi_{Nj} & \cdot & \pi_{NN} \end{pmatrix} = ((\pi_{ij}))$$

A design $p^* \in P_n$ will be said to be an inclusion probability proportional to size (π^*ps) sampling design if it realizes pre assigned values of the first order inclusion probabilities $\pi_i^* = \sum_{s \supset i} p^*(s)$ with $\sum_i \pi_i^* = n$. Such a π^*ps design can be constructed in various ways. Details are given by Brewer and Hanif (1988). The class of π^*ps design will be denoted by P_n^*. Clearly $P_n^* \subset P_n$.

Gabler (1987) considered a situation where a statistician would like to implement a sampling design $p_0 (\in P_n)$ with a given inclusion probability matrix $\pi_0 = \left((\pi_{ij}^0)\right)$, but out of theoretical consideration, a $p^* (\in P_n^*)$ design is desirable. He therefore, recommended a design $p_1^* (\in P_n^*)$ called a "nearest π^*ps design" which is as near as possible to p_0 in the sense of minimizing the distance.

$$D(p_0, p_1^*) = \sum_{s \in T(p_0)} \frac{[p_1^*(s) - p_0(s)]^2}{p_0(s)}$$

Gabler showed that $D(p_0, p_1^*)$ attains a minimum value when

$$p_1^*(s) = p_0(s) \sum_{i \in s} \lambda_i \quad \forall\, s \in S \tag{1}$$

where $\lambda' = (\lambda_1, \ldots, \lambda_i, \ldots, \lambda_n)$ satisfies the equation

$$\pi_0 \lambda = \pi^* \tag{2}$$

and π^* = Transpose of $(\pi_1^*, \ldots, \pi_i^*, \ldots, \pi_N^*)$

Clearly, $p_1^*(s)$ in (1) can take a negative value since $D(p_0, p_1^*)$ is minimized ignoring the restriction $p_1^*(s) > 0$. Gabler (1987) used the equation (1) in constructing a rejective π^*ps sampling plan which is described as follows. Suppose we want to implement a π^*ps sampling design p for the efficiency point of view. But, for practical considerations, a set of undesirable samples S_0 with $p(s) > 0$ for $s \in S_0$ is excluded from S and a design p_0^* is constructed by assigning selection probability as

$$p_0^*(s) = \begin{cases} \dfrac{p(s)}{\sum\limits_{s' \notin S_0} p(s')} & \text{for } s \notin S_0 \\ 0 & \text{for } s \in S_0 \end{cases} \quad (3)$$

Obviously the design p_0^* defined in (3) is generally not a π^*ps design. So, we look for a π^*ps design as near as possible to p_0^* in Gabler's (1987) sense. Certainly, such a design in obtained from the equations (1) and (2) whenever $\sum\limits_{i \in s} \lambda_i$ is positive $\forall s \in S$ and design is given by

$$\tilde{p}(s) = p_0^*(s)\left(\sum_{i \in s} \lambda_i\right) \text{for } s \in S - S_0 \text{ with } \pi_0^* \lambda = \pi^* \quad (4)$$

where π_0^* = inclusion probability matrix of the design p_0^*.

Adhikary (1996) studied conditions under which the system of non-homogeneous linear equations (4) are consistent and admit nonnegative solution of λ and derived a set of sufficient conditions based on the following theorems.

Theorem 1.1: If $i \notin S_0$, then $\pi_0^* \lambda = \pi^*$ does not possess a nonnegative solution for λ.

Theorem 1.2: If all units are evenly distributed over S_0 $(S - S_0)$, then rank $\pi_0^* = N$.

Theorem 1.3: If all the units are evenly distributed over S_0, then $\pi_0^* \lambda = \pi^*$ admits a nonnegative solution for λ.

In this article, it is shown that all the theorems mentioned above are incorrect. In addition, some new results related to the existence of nonnegative solution of nonnegative solution of λ are presented.

2. Results

2.1 Disproval of Adikary's (1996) Assertions

2.1.1 Theorem 1.1

Consider $U = \{1, 2, 3, 4, 5\}$, $N = 5$, $n = 2$, normed size measures p_i's are $p_1 = .275$, $p_2 = .175$, $p_3 = .225$, $p_4 = .175$, $p_5 = .15$ and a design p^* with $\pi_i^* = np_i$ i.e. $\pi_1^* = .55$, $\pi_2^* = .35$, $\pi_3^* = .45$, $\pi_4^* = .35$, $\pi_5^* = .30$ as follows

s	(1,2)	(1,3)	(1,4)	(1,5)	(2,3)	(2,4)	(2,5)	(3,4)	(3,5)	(4,5)
$p^*(s)$	.10	.15	.15	.15	.15	.05	.05	.10	.05	.05

Here the support of $p^* = T^* = \{(1, 2), (1, 3), (1, 4), (1,5), (2,3), (2,4), (2, 5), (3, 4), (3, 5), (4,5)\}$

Let the undesirable sample, $S_0 = \{(1,2), (3,4)\}$. Note that the unit $5 \notin S_0$ and p_0^* is given by

s :	(1,3)	(1,4)	(1,5)	(2,3)	(2,4)	(2,5)	(3,5)	(4,5)
$p_0^*(s) = p^*(s)/.8$:	.1875	.1875	.1875	.1875	.0625	.0625	.0625	.0625

The inclusion probability matrix for the design p_0^* is given by

$$\pi_0^* = \begin{pmatrix} .5625 & 0 & .1875 & .1875 & .1875 \\ & .3125 & .1875 & .0625 & .0625 \\ & & .4375 & 0 & .0625 \\ & & & .3125 & .0625 \\ & & & & .3750 \end{pmatrix}$$

Now the equation

$$\pi_0^*\lambda = \pi^* = \begin{pmatrix} .55 \\ .35 \\ .45 \\ .35 \\ .30 \end{pmatrix}$$

gives

$\lambda_1 = .525356$, $\lambda_2 = .652991$, $\lambda_3 = .488889$, $\lambda_4 = .625641$ and $\lambda_5 = .242735$

Here all λ_i's are positive which contradicts to the Theorem 1.1. Hence Theorem 1.1 is false.

Remark 2.1. It can be pointed out that the incorrect step of the Adhikary's **Theorem 1.1** is as follows. In the third line of his Remark 3.1.III (Adhikary (1996), page 1763) the author attempts to show that the inequality $np_i\pi_j^0 < \pi_{ij}^0$ is consistent for all $i \neq j$ by summing both sides of the inequality over all $j \in U = \{1, \ldots, N\}$. Hence Adhikary's **Theorem 1.1** may be corrected as follows

Corrected version of the Theorem 1.1 : A sufficient condition for the nonexistence of a non-negative solution for λ in the system $\pi_0\lambda = \pi^*$ is that $\pi_{ij}^0 > np_i\pi_j^0, \forall\, i \neq j \in U$.

2.1.2. Theorem 1.2

Theorem 1.2 is incorrect since it is based on the wrong argument presented in the equation (3.7) of Adhikary's (1996) paper. However, we can check it through the following example.

Let $U = \{1, 2, 3, 4\}$, $N = 4$, $n = 2$, $p_1 = .2$, $p_2 = p_3 = .25$, $p_4 = .3$

The design p^* with $\pi^* = \begin{pmatrix} .4 \\ .5 \\ .5 \\ .6 \end{pmatrix}$ is as follows

s	(1, 2)	(1, 3)	(1, 4)	(2, 3)	(2, 4)	(3, 4)
$p^*(s)$	.05	.15	.20	.20	.25	.15

$T^* = \{(1, 2), (1, 3), (1, 4), (2, 3), (2, 4), (3, 4)\}$

Let the undesirable sample be (1, 2) and (3, 4) i.e. $S_0 = \{(1, 2), (3, 4)\}$

The design p_0^* is given by

s	(1, 3)	(1, 4)	(2, 3)	(2, 4)
$p_0^*(s) = p^*(s)/.8$	.1875	.25	.25	.3125

Here every unit is repeated only once in S_0, i.e. the units are evenly distributed over S_0. According to the Theorem 1.2, the rank of the inclusion probability matrix

$$\pi_0^* = \begin{pmatrix} .4375 & 0 & .1875 & .2500 \\ & .5625 & .2500 & .3125 \\ & & .4375 & 0 \\ & & & .5625 \end{pmatrix}$$

of the design p_0^* should be of full rank 4. But we note that rank of π_0^* is less than 4 since the sum of the first two rows of the matrix π_0^* is equal to the sum of the last two rows of π_0^*.

2.1.3. Theorem 1.3

Theorem 1.3 is also incorrect because its proof is based on the result of the Theorem 1.2 which is incorrect. For the sake of clarity, it can be checked that the system of equations $\pi_0^* \lambda = \pi^*$ (with π_0^* and π^* given in the Section 2.1.2) do not admit any solution for λ since the rank of $\pi_0^* = 3$ while the rank of $(\pi_0^*, \pi^*) = 4$.

2.2 Some Additional Results

2.2.1 Existence of Positive Solution for λ

Theorem 2.1. A necessary condition for positive solution of λ for the equation (2) is that both the designs p_0 and p_1^* must have the same support i.e. $T(p_0) = T(p_1^*)$.

Proof : Let π_i^* be the inclusion probability of the i^{th} unit of the sampling design p_1^*. i.e.

$$\pi_i^* = \sum_{s \in T(p_1^*)} I_{si} p_1^*(s) = \sum_{s \in T(p_1^*)} I_{si}\left(\sum_{j \in s} \lambda_j\right) p_0(s) = \sum_{s \in T(p_1^*)} I_{si}\left(\sum_{j} \lambda_j I_{sj}\right) p_0(s)$$

Now suppose $T(p_1^*)$ is contained in $T(p_0)$ (as assumed by Gabler (1987)), then we have

$$\pi_i^* = \sum_{s \in T(P_0)} I_{si}\left(\sum_{j} \lambda_j I_{sj}\right) p_0(s) - \sum_{s \in T^c} I_{si}\left(\sum_{j} \lambda_j I_{sj}\right) p_0(s) \text{ with } T^c = T(p_0) - T(p_1^*)$$

$$= \sum_{j} \lambda_j \pi_{ij}^0 - \sum_{j} \lambda_j \sum_{s \in T^c} p_0(s) I_{si} I_{sj}$$

i.e. $\sum_{j} \lambda_j \sum_{s \in T^c} p_0(s) I_{si} I_{sj} = 0$ (5)

$$\left(\text{since} \sum_{j} \lambda_j \pi_{ij}^0 = \pi_i^*\right)$$

From (5), we note that all λ_j's $(j = 1, \ldots, N)$ cannot be positive unless $T(p_0) = T(p_1^*)$ as $\sum_{j} \lambda_j \sum_{s \in T^c} p_0(s) I_{si} I_{sj} \geq 0$, for every i and j. This proves the theorem.

2.2.2. Existence of unique solution of λ

The system of equations $\pi_0\lambda = \pi^*$ given in (2) has a unique solution when π_0 is full rank N. Let us search for the conditions under which π_0 becomes full rank. Suppose that the designs p_1^* and p_0 have the same support i.e. $T(p_1^*) = T(p_0) = (s_1, \ldots, s_r \ldots, s_b)$, where b = total number of samples belonging to $T(p_1^*)$ or $T(p_0)$. Consider the following matrix D of order $N \times b$

$$D = \begin{pmatrix} I_1(s_1) & \ldots & I_1(s_j) & \ldots & I_1(s_b) \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ I_k(s_1) & \ldots & I_k(s_j) & \ldots & I_k(s_b) \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ I_N(s_1) & \ldots & I_N(s_j) & \ldots & I_N(s_b) \end{pmatrix}$$

where the elements $I_k(S_r) = 1$ if $k \in s_r$ and zero otherwise; $r = 1, \ldots, b$; $k = 1, \ldots, N$.

Let us consider a unit and a sample of a sampling design as a treatment and a block of an incomplete block design. Writing

$$\pi_{ij}^0 = \sum_{s \in T(p)} I_{si} I_{sj} p_0(s) = \sum_{r=1}^{b} I_i(s_r) I_j(s_r) p_0(s) = D_i P D_j^T$$

where D_j^T = Transpose of D_j (D_j = the jth row of D)

$$P = \begin{pmatrix} p_0(s_1) & \ldots & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & p_0(s_r) & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & & 0 & & p_0(s_b) \end{pmatrix}$$

we get

$$\pi_0 = D P D^T$$

Now, the rank of π_0 = Rank of $D P D^T$ = Rank $D^T D P$

= Rank of $D^T D$ = Rank of DD^T

and

$$DD^T = \begin{pmatrix} r_{11} & \cdots & r_{1j} & \cdots & r_{1N} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ r_{i1} & \cdots & r_{ij} & \cdots & r_{iN} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ r_{N1} & & r_{Nj} & & r_{NN} \end{pmatrix}$$

where

$$r_{ii} = \sum_{i=1}^{b} I_i(s_r) = \text{ total number of samples containing the } i^{th} \text{ unit}$$

$$r_{ij} = \sum_{i=1}^{b} I_i(s_r) I_j(s_r) = \text{total number of samples containing the } i^{th} \text{ and } j^{th}\ (i \neq j) \text{ unit}$$

$$\leq r_{ii}$$

Now if $r_{ii} = r$ for every i and $r_{ij} = \mu\,(< r)$, for all $i \neq j$, then D corresponds to the incidence matrix of a balanced incomplete block design (BIBD) with parameters b = total number of samples (blocks), N = total number of units (treatments) = v, n = sample (block) size = k, r = replication of the ith unit (treatment) in all samples (blocks) and $\mu =$ total number of times of the appearance of any two units (treatments) in the same sample (block). Furthermore, the determinant of $D\,D^T$ is nonnegative since $r > \mu$ and hence the matrix is of full rank N. Thus, we prove the following theorem.

Theorem 2.2. If the incidence matrix D, of a sampling design p_0 corresponds to the incidence matrix of a balanced incomplete block design, then the rank of $\pi_0 = N$ and the system $\pi_0\,\lambda = \pi^*$ given in (2) has the unique solution, $\lambda = \pi_0^{-1}\,\pi^*$.

From the Theorem 2.2, we note that when a sampling design p_0 with support $T_{(p_0)}$ comprises $\binom{N}{n}$ possible samples of size n, the incidence matrix D of the sampling design p_0 corresponds to an incidence matrix of a BIBD with parameters $b = \binom{N}{n}$, $v = N$, $r = \binom{N-1}{n-1}$, $k = n$ and $\mu = \binom{N-2}{n-2}$. This gives the following result.

Theorem 2.3. If the support of a sampling design p_0 consists of $\binom{N}{n}$ possible samples of size n, the rank of π_0 is N and the equation (2) has the unique solution, $\lambda = \pi_0^{-1}\pi^*$.

Remark 2.2. Except for systematic sampling designs, the designs based on without replacement sampling schemes which are most often used in practice have support of $\binom{N}{n}$ samples and hence provide a unique solution for λ.

2.2.3. Existence of Gabler's (1987) design

We define a sampling design $p^*(s) = \left(\sum_{i \in s} \lambda_i\right) p_0(s)$ satisfying $\pi_0\lambda = \pi^*$ with $\left(\sum_{i \in s} \lambda_i\right) \geq 0$, as Gabler's design. Existence of such a design obviously depends on the structure of the inclusion probability matrix π_0. We will now consider the existence of Gabler's design for a few typical types of the matrix π_0. We see that the equations (1) and (2) yield

$$\sum_i \lambda_i \pi_i^0 = 1 \tag{6}$$

Example 2.1. Let us suppose that $\pi_i^0 = c, \forall\, i$ and $\pi_{ij}^0 = d$ for $i \neq j$. Then $\sum_i \lambda_i \pi_{ij}^0 = \pi_i^*$ gives $(c - d)\lambda_i + d\sum_i \lambda_i = \pi_i^*$ which in turn yields $\lambda_i = \dfrac{\left(\pi_i^* - \frac{c}{d}\right)}{(c - d)}$, since $\sum_i \lambda_i = \frac{1}{c}$ and $p^*(s) = \left(\sum_{i \in s} \lambda_i\right) p_0(s)$ is nonnegative whenever $\sum_{i \in s} \pi_i^* \geq n\frac{d}{c}$, where each of the sample s consists of $n\,(< N)$ distinct units.

Remark 2.3. When p_0 is an SRSWOR sampling design, we get $\pi_i^0 = \frac{n}{N}$ and $\pi_{ij}^0 = \dfrac{n(n-1)}{\{N(N-1)\}}$ and the Gabler's $(c - d)\lambda_i + d\sum_i \lambda_i = \pi_i^*$ which in turn yields $\lambda_i = \dfrac{\left(\pi_i^* - \dfrac{d}{c}\right)}{(c-d)}$, since $\sum_i \lambda_i = \dfrac{1}{c}$.

Solution exists whenever $\sum_{i \in s} \pi_i^* \ge n\dfrac{n-1}{N-1}$ holds for all possible $M = \binom{N}{n}$ samples. (7)

Remark 2.4. Suppose that we have a situation where the above condition (7) does not hold for all possible M samples but does hold for a fewer set of samples S_1 that can form a BIBD with parameters b, v, r, k and μ (defined in Section 2.2.2). In such a situation we can construct a sampling design realizing the same set of preassigned inclusion probabilities p_i^*'s through a design p_0 with support $T(p_0) = S_1$, by assigning equal probability $1/b$ to each of the sample. Since, in this case $\pi_i^0 = \dfrac{r}{b}, \pi_{ij}^0 = \dfrac{\mu}{b}$, $\lambda_i = b(\pi_i^* - \mu / r)(r - \mu)$ and $\sum_i \lambda_i \ge 0$ whenever $\sum_{i \in s} \pi_i^* \ge n\dfrac{\mu}{r} = \dfrac{n(n-1)}{(N-1)}$ [noting the parameters of BIB design satisfy $\mu(v-1) = r(k-1)$, i.e. $\mu(N-1) = r(n-1)$, for details see Raghvarao (1971)].

Example 2.2. Poisson sampling scheme

Here $\pi_{ij}^0 = \pi_i^0 \pi_j^0$ for $i \ne j$; and $\pi_i^* = \sum_j \pi_{ij}^0 \lambda_j$ gives $\lambda_i = \dfrac{1}{1-\pi_i^0}\left(\dfrac{\pi_i^*}{\pi_i^0} - k\right)$

where $k = \dfrac{\sum_{j=1}^{N} \dfrac{\pi_j^*}{1-\pi_j^0}}{1 + \sum_{j=1}^{N} \dfrac{\pi_j^0}{1-\pi_j^0}}$

A sufficient condition of existence of Gabler's design is $\sum_{i \in s} \lambda_i = \sum_{i \in s} \dfrac{1}{1-\pi_i^0}\left(\dfrac{\pi_i^*}{\pi_i^0} - k\right) \ge 0$

For a Poisson sampling scheme, the sample size is not fixed. Consider a sample $s_1 = \{i\}$ containing just one unit. In this case for the existence of Gabler's design we must have

$\sum_{i \in s_1} \lambda_i = \lambda_i = \frac{1}{1-\pi_i^0}\left(\frac{\pi_i^*}{\pi_i^0} - k\right) \geq 0$ for every i. So, Gabler's design exists if and only if $\pi_i^* \geq k\pi_i^0$ for every i. It might be worth noting that the condition $\pi_i^* \geq k\pi_i^0, \forall\, i \in U$ implies that the expected sample size of the new design p^* would be at least k times as large as the sample size of the original design p_0.

REFERENCE

Adhikary, A. (1996). On non-negativity of the nearest proportional to size sampling design. *Comm. Stat.–Theory Methods,* **25,** 1757–1768.

Brewer, K.R.W. and Hanif, M. (1983). *Sampling with Unequal Probabilities.* Springer-Verlag, New York.

Gabler, S. (1987). The nearest proportional to size sampling design. *Comm. Stat.–Theory Methods,* **16,** 1117–1131.

Midzuno, H. (1952). On the sampling system with probability proportional to sum of sizes. *Ann. Inst. Statist. Maths.,* **3,** 99–107.

Raghvarao, D. (1971). *Constructions and Combinatorial Problems in Designs of Experiments.* Wiley, New York.

Sen, A.R. (1953). On the estimation of variance in sampling with varying probabilities. *J. Ind. Soc. Agril. Statist.,* **5,** 119-127.