The Annals of Statistics
1981, Vol. 9, No. 1, 229–232

ON THE COMPLETENESS OF THE CLASS OF FIXED SIZE SAMPLING STRATEGIES

BY H. STENGER AND S. GABLER

University of Mannheim

A sampling strategy consists of a sampling design and an estimator. Of special importance are strategies combining simple random sampling with the sample mean as an estimator. We consider this class of strategies and the smaller class of strategies with simple random sampling of (almost) fixed sample size.

The convexity of the underlying loss function does not imply the completeness of the smaller class. However the strict convexity of the loss function together with the convexity of its second derivative, is sufficient for this completeness.

1. Introduction. Consider the class $\mathscr{T}$ of all sampling strategies (p, e_0) where p denotes a simple random sampling procedure and e_0 the sample mean. The restriction to $\mathscr{T}$ can be justified by a modification of the minimax principle (see Wesler 1959). The only condition needed is the convexity of the loss function (see Stenger 1979a). If the usual quadratic loss function is adopted, the subclass $\mathscr{T}_0$ of all strategies, (p, e_0) where p is of fixed size or nearly so, in a sense to be defined, is complete in $\mathscr{T}$ (see Ramakrishnan 1969). We are interested in general conditions sufficient for the completeness of $\mathscr{T}_0$ in $\mathscr{T}$.

2. General definitions and notations. Consider a finite population consisting of N distinguishable units labelled $i = 1, 2, \cdots, N$. With each i is associated an unknown variate value x_i. X is the set of all possible parameters $\mathbf{x} = (x_1, x_2, \cdots x_N)$.

S is the set of all samples, i.e., the set of all nonempty subsets of $\{1, 2, \cdots N\}$. The size $n(s)$ of a sample $s \in S$ is the number of elements in s.

A function p on S is said to be a sampling design if

$$p(s) \geqq 0 \qquad \text{for all } s \in S,$$
$$\textstyle\sum_{s\in S} p(s) = 1.$$

We denote

$$\textstyle\sum_{s\in S} n(s)\, p(s)$$

as the average sample size of p.

A function e: $S \times X \rightarrow \mathbb{R}$ depending on $\mathbf{x} \in X$ only through x_i $(i \in s)$ is called an estimator. Of special importance as an estimator is the sample mean

$$e_0(s, \mathbf{x}) = \frac{1}{n(s)} \textstyle\sum_{i\in s} x_i.$$

A strategy consists of a design p and an estimator e. If $\mathbf{x} \in X$ is the true parameter and a is our estimate, a loss $L(\mathbf{x}, a)$ arises. The risk function R is defined as

$$\mathrm{R}(\mathbf{x}; p, e) = \textstyle\sum_{s\in S} [L(\mathbf{x}, e(s, \mathbf{x})) + cn(s)]p(s).$$

In this formula c is a positive constant denoting the cost of drawing one element from $\{1, 2, \cdots N\}$.

Received December 1979.

AMS 1970 *subject classifications.* Primary 62 D 05; secondary 62 C 07.

Key words and phrases. Finite population, simple random sampling and symmetric estimation, complete classes, fixed size sampling strategies, loss functions.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Statistics.* www.jstor.org

The strategy (p', e') is better than the strategy (p, e) if

$$R(\mathbf{x}; p', e') \leqq R(\mathbf{x}, p, e) \qquad \text{for all} \quad \mathbf{x} \in X$$

with strict inequality for at least one $\mathbf{x} \in X$.

Let C and C_0 be classes of strategies with $C_0 \subset C$. C_0 is said to be complete in C if for each strategy $(p, e) \in C \cap \bar{C}_0$ there exists a strategy $(p', e') \in C_0$ which is better than (p, e).

3. Simple random sampling of almost fixed size. For $m = 1, 2, \cdots N$ we define

$$\begin{aligned} p_m(s) &= \frac{1}{\binom{N}{m}} && \text{for } s \in S \text{ with } n(s) = m, \\ &= 0 && \text{for } s \in S \text{ with } n(s) \neq m. \end{aligned}$$

Then p_m is a sampling design.

A sampling design p is called symmetric if it is a probability mixture of $p_1, p_2, \cdots p_N$. Symmetry of p is therefore equivalent to the existence of a probability vector $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \cdots \alpha_N)$ with

$$p(s) = \textstyle\sum_{i=1}^N \alpha_i p_i(s) \qquad \text{for all } s \in S.$$

The sampling design $p = \sum \alpha_i p_i$ has the average sample size

$$\textstyle\sum_{i=1}^N i\alpha_i.$$

We associate with $\boldsymbol{\alpha}$ the probability vector $\boldsymbol{\alpha}^* = (\alpha_1^*, \cdots \alpha_N^*)$ by defining

$$\begin{aligned} \alpha_i^* &= 1 - (\textstyle\sum j\alpha_j - [\sum j\alpha_j]) && \text{for } i = [\textstyle\sum j\alpha_j], \\ &= \textstyle\sum j\alpha_j - [\sum j\alpha_j] && \text{for } i = [\textstyle\sum j\alpha_j] + 1, \\ &= 0 && \text{otherwise,} \end{aligned}$$

where $[z]$ denotes the largest integer not exceeding z. It is easy to see that the symmetric sampling design $p^* = \sum \alpha_i^* p_i$ has the same average sample size $\sum i\alpha_i$ as p, and that $p^*(s) > 0$, $p^*(s') > 0$ implies $| n(s) - n(s') | \leqq 1$. If $\sum i\alpha_i$ is an integer, m, we have $p^* = p_m$. Otherwise, p^* is a probability mixture of $p_{[\sum i\alpha_i]}$ and $p_{[\sum i\alpha_i]+1}$. When the sampling design p^* is used, the sample size $n(s)$ does not vary more than is necessary to realize the average sample size $\sum i\alpha_i$ of p.

We denote a symmetric sampling design p as simple random sampling of almost fixed size if $p^* = p$.

4. A completeness theorem. Let us define

$$\mathscr{T} = \{(p, e_0) : p \text{ symmetric}\},$$

$$\mathscr{T}_0 = \{(p, e_0) : p^* = p\}.$$

If

$$L(\mathbf{x}, a) = l_0(\mathbf{x})(\bar{x} - a)^2$$

with l_0 strictly positive and $\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$ it can be shown (see Ramakrishnan 1969) that $\mathscr{T}_0$ is complete in $\mathscr{T}$. For the quadratic loss function above, simple random sampling of almost fixed size should therefore be used if e_0 has been accepted as an estimator and if only symmetric designs are acceptable.

One might suppose that the same result holds for all loss functions $L(\mathbf{x}, a)$ convex in a. The following example shows that this is not the case. For $\mathbf{x} \in X = \mathbb{R}^4$, let

$$L(\mathbf{x}, a) = |\bar{x} - a|.$$

Then we have for $\mathbf{x} = (-\epsilon, -\epsilon, \epsilon, \epsilon)$ with $\epsilon > 0$

$$R(\mathbf{x}; p_1, e_0) = \epsilon + c,$$

$$R(\mathbf{x}; p_2, e_0) = \frac{\epsilon}{3} + 2c,$$

$$R(\mathbf{x}; p_3, e_0) = \frac{\epsilon}{3} + 3c,$$

$$R(\mathbf{x}; p_4, e_0) = 4c.$$

The class $\mathscr{T}_0$ is given by

$$\{\lambda_i p_i + (1 - \lambda_i)p_{i+1} : 0 \leq \lambda_i \leq 1, i = 1, 2, 3\}.$$

Consider the symmetric design $\bar{p} = ½ (p_2 + p_4)$. Clearly, $\bar{p} \notin \mathscr{T}_0$ and

$$R(\mathbf{x}; \bar{p}, e_0) < R(\mathbf{x}; \lambda_i p_i + (1 - \lambda_i) p_{i+1}, e_0) \qquad \text{for } i = 1, 2$$

if $\epsilon > 6c$, and

$$R(\mathbf{x}; \bar{p}, e_0) < R(\mathbf{x}; \lambda_3 p_3 + (1 - \lambda_3)p_4, e_0)$$

if $\epsilon < 6c$. Thus, $\mathscr{T}_0$ is not complete in $\mathscr{T}$.

We now give conditions sufficient for the completeness of $\mathscr{T}_0$ in $\mathscr{T}$.

THEOREM. *Let $L(\mathbf{x}, a)$ be strictly convex in $a \in I$, I an open interval. Let the second partial derivative $\partial^2 L(\mathbf{x}, a)/\partial a^2$ of the loss function L exist for $a \in I$. Suppose further that this derivative is convex and that $X \subset I^N$. Then $\mathscr{T}_0$ is complete in $\mathscr{T}$.*

REMARK 1. Let X be any subset of $\mathbb{R}^N$ and

$$L(\mathbf{x}, a) = 1_0 (\mathbf{x})(\bar{x} - a)^{2k} \qquad k \in N$$

with 1_0 strictly positive. Clearly, $L(\mathbf{x}, a)$ is strictly convex and

$$\frac{\partial^2 L(\mathbf{x}, a)}{\partial a^2} = 1_0 (\mathbf{x})\, 2k(2k - 1)\, (\bar{x} - a)^{2k-2}$$

is convex. Then the theorem implies the completeness of $\mathscr{T}_0$ in $\mathscr{T}$.

REMARK 2. Let X be any subset of $\mathbb{R}^N_+ = \{\mathbf{x} \in \mathbb{R}^N : x_i > 0\}$. For this case Stenger (1979b) proposed to use the loss function

$$L(\mathbf{x}, a) = 1_0(\mathbf{x}) \left(\frac{a}{\bar{x}} - 1 - \log \frac{a}{\bar{x}}\right)$$

with $1_0(\mathbf{x})$ strictly positive. As

$$\frac{\partial^2 L(\mathbf{x}, a)}{\partial a^2} = 1_0(\mathbf{x})/a^2$$

is positive and convex in a the completeness of $\mathscr{T}_0$ in $\mathscr{T}$ follows.

Furthermore, we have for all $\mathbf{x} \in X$ and $\lambda \neq 1$

$$R(\mathbf{x}; p, e_0) < R(\mathbf{x}; p, \lambda e_0\}$$

if p is any symmetric design (see Stenger (1979b)). $\mathscr{T}_0$ therefore is complete even in the class

$$\mathscr{T}' = \{(p, \lambda e_0) : p \text{ symmetric}, \lambda > 0\}.$$

5. Proof of the theorem. We denote the real sequence $h_1, h_2, \ldots, h_N$ as strictly convex if for $k = 2, \ldots, N - 1$

$$2h_k < h_{k-1} + h_{k+1}.$$

LEMMA 1. *Let $h_1, h_2, \ldots, h_N$ be a strictly convex sequence and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \cdots \alpha_N)$ a probability vector. Then*

$$\sum_1^N \alpha_j h_j \geq \sum_1^N \alpha_j^* h_j. \tag{1}$$

Equality holds if and only if $\boldsymbol{\alpha} = \boldsymbol{\alpha}^*$.

PROOF OF LEMMA 1. $\boldsymbol{\alpha}^* = \boldsymbol{\alpha}$ clearly implies equality in (1). If $\alpha^* \neq \alpha$, we have $\alpha_N < 1$ and $i_0 = [\sum_1^N j\alpha_j] \in \{1, 2, \cdots N-1\}$. As $h_1, h_2, \ldots, h_N$ is a strictly convex sequence, we derive

$$h_i > h_{i_0} + (h_{i_0+1} - h_{i_0})(i - i_0) \tag{2}$$

for $i \neq i_0, i_0 + 1$. Multiplication of both sides of (2) by α_i and the subsequent summation over i yields

$$\sum_1^N \alpha_j h_j > \sum_1^N \alpha_j^* h_j.$$

LEMMA 2. *Let (a, b) be an open interval and $f\colon (a, b) \to \mathbb{R}$ a strictly convex function with f'' convex. Then*

$$f_i = \frac{1}{\binom{N}{i}} \sum_{s:n(s)=i} f(e_0(s, \mathbf{x})); \qquad i = 1, 2, \cdots, N$$

is a strictly convex sequence for all $x_1, x_2, \ldots, x_N$ with $x_i \in (a, b)$ and for all integers $N \geq 3$ unless $x_1 = x_2 = \cdots = x_N$.

The proof of lemma 2 is given by Gabler (1979).

We now prove the theorem. For $p = \sum_1^N \alpha_i p_i$ and $\mathbf{x} \in X$ with $x_1 = x_2 = \cdots = x_N$ we clearly have

$$R(\mathbf{x}; p, e_0) = R(\mathbf{x}; p^*, e_0).$$

If $\mathbf{x} \in X$ implies $x_1 = x_2 = \cdots = x_N$, the theorem is true as the strategy $(p_1, e_0) \in \mathscr{T}_0$ is better than any other strategy in $\mathscr{T}$. Now let $\mathbf{x}$ be an element of X for which $x_1 = x_2 = \cdots = x_N$ is not true and let L be a loss function satisfying the conditions of our theorem. From Lemma 2 it follows that

$$l_i = \frac{1}{\binom{N}{i}} \sum_{s:n(s)=i} L(\mathbf{x}, e_0(s, \mathbf{x})); \qquad i = 1, 2, \cdots, N$$

is a strictly convex sequence. We conclude therefore

$$\sum_1^N \alpha_i \mathrm{l}_i > \sum_1^N \alpha_i^* \mathrm{l}_i$$

for $p = \sum_1^N \alpha_i p_i$ with $p^* \neq p$. The last inequality yields

$$R(\mathbf{x}; p, e_0) > R(\mathbf{x}, p^*, e_0)$$

and the theorem is proved.

REFERENCES

[1] GABLER, S. (1979). Folgenkonvexe Funktionen. *Manuscripta Math.* **29** 29–47.

[2] RAMAKRISHNAN, M. K. (1969). Some results on the comparison of sampling with and without replacement. *Sankhyā Ser. A* **31** 333–342.

[3] STENGER, H. (1979a). A minimax approach to randomization and estimation in survey sampling. *Ann. Math. Statist.* **7** 395–399.

[4] STENGER, H. (1979b). Loss functions and admissible estimators in survey sampling. *Metrika* **26** 205–214.

[5] WESLER, O. (1959). Invariance theory and a modified minimax principle. *Ann. Math. Statist.* **30** 1–20.

UNIVERSITY OF MANNHEIM
68 MANNHEIM, GERMANY