The Annals of Mathematical Statistics
1972, Vol. 43, No. 4, 1323-1328

A NOTE ON HYPERADMISSIBILITY OF ESTIMATORS FOR FINITE POPULATIONS

BY V. M. JOSHI

Secretary, Maharashtra Government, Bombay

A condition on the sampling design which previously [2] was shown to be a sufficient condition for the Horvitz-Thompson estimator to be the unique hyperadmissible estimator of the population total is now shown to be a necessary condition also save for trivial exceptions. The exceptions are the sampling designs in which the only samples with positive probability are such that they include the whole population.

1. Introduction. In a previous paper [2], it was shown that the Horvitz-Thompson estimator (H-T estimator for short) for the population total is the unique hyperadmissible estimator for the population total, if the sampling design satisfies a certain mild condition. The question whether this condition is necessary also, for the uniqueness of the H-T estimator has remained open. This is investigated in the following, and it is shown that the condition is necessary also except for certain trivial sampling designs. The excepted sampling designs are those in which the only samples with positive probability are such that they include the whole population. In the course of the demonstration, we obtain a simpler form of the condition on the sampling design, which is easier to apply to any given design.

2. Preliminaries. We use the same notation and definitions as in [2]. For convenience the relevant notation is reproduced below. $\mathcal{U}$ denotes a finite population of units U_i, $i = 1, 2, \cdots, N$. A sample s means any finite, ordered sequence of units, repetitions being allowed. S denotes the set of all possible samples s. A sampling design is obtained by defining on S a probability measure P. For a given sampling design, P_s denotes the probability of the sample s. With each unit U_i, $i = 1, 2, \cdots, N$, is associated a real number Y_i. The vector $\mathbf{Y} = (Y_1, Y_2, \cdots, Y_N)$ denotes a point in the Euclidean space R_N. The population total is $Y = \sum_{i=1}^{N} Y_i$. An estimator T is a function defined on $S \times R_N$, which for each $s \in S$, depends on $\mathbf{Y}$, through only those Y_i, for which the unit U_i occurs in the sample (sequence) s. Unbiasedness and admissibility with respect to a loss function of an estimator T of the population total are defined as usual. Here 'admissible' means 'admissible within the class of unbiased estimators'. An estimator is hyperadmissible, if it is unbiased and admissible, and is also admissible in every co-ordinate subspace $R(i_1, i_2, \cdots, i_m)$, defined by $1 \leqq i_1, i_2, \cdots, i_m \leqq N$, $m \geqq 1$; $Y_j \neq 0$ if $j \in [i_1, i_2, \cdots, i_m]$ and $Y_j = 0$ otherwise. π_i, $i = 1, 2, \cdots, N$, denotes the inclusion probability of the unit U_i, i.e. the total probability (in a given sampling design) of all samples in which the unit

Received May, 25, 1971; revised October 29, 1971.

U_i occurs at least once. Similarly π_{ij} denotes the joint inclusion probability of the pair of units U_i and U_j. We consider only sampling designs such that

$$\pi_i > 0\,, \qquad i = 1, 2, \cdots, N\,. \tag{1}$$

This restriction is necessary, as otherwise unbiased estimation of the population total is not possible. $\mathscr{D}_0$ denotes the class of sampling designs such that the only samples with positive probability $P_s > 0$ are such that all the population units occur in them.

The H-T estimator is defined by

$$\hat{Y}_{HT} = \sum_{i \in s} \frac{Y_i}{\pi_i} \tag{2}$$

where in the right-hand side the sum is taken over all units U_i which occur in the sample s, each unit being taken once only, irrespective of the number of times it occurs in the sample.

For a given sampling design, we denote by $\bar{S}$, the subset of S, consisting of all the samples s for which $P_s > 0$ and by $\bar{S}^*(i)$, $i = 1, 2, \cdots, N$ the subset of $\bar{S}$ consisting of all the samples $s \in \bar{S}$, in which the unit U_i does not occur.

3. Main result. The uniqueness of the H-T estimator as a hyperadmissible estimator for the population total was proved in [2], subject to the sampling design satisfying the following condition:

CONDITION C_1. There exists an ordered sequence of integers $i_1, i_2, \cdots, i_k$, such that for each m, $1 \leqq m \leqq k-1$, the sets $\bigcup_{r=1}^{r=m} \bar{S}^*(i_r)$ and $\bar{S}^*(i_{m+1})$ have at least one sample in common and

$$\bigcup_{r=1}^{r=k} \bar{S}^*(i_r) = \bar{S}\,. \tag{3}$$

We shall show that the condition C_1 is necessary also for the uniqueness of the H-T estimator, except for sampling designs of the class $\mathscr{D}_0$. We first formulate an alternative condition on the sampling design.

CONDITION C_2. The sampling design is not of the class $\mathscr{D}_0$ and moreover is such that it is possible to partition the set $\bar{S}$ of all samples with $P_s > 0$, into non-empty subsets G_1 and G_2 and the population $\mathscr{U}$ into sets of units g_1 and g_2, one of which may in special cases be empty, and which are such that

(i) all the units in g_1 appear in each sample $s \in G_1$,
(ii) all the units in g_2 appear in each sample $s \in G_2$.

We shall next prove the following.

PROPOSITION 3.1. *A sampling design satisfies the condition* C_2, *if and only if, it does not satisfy* C_1 *and is not the class* $\mathscr{D}_0$.

PROOF. We shall first prove the 'if' part of the proposition. We assume accordingly that C_1 does not hold and the sampling design is not of the class

$\mathscr{D}_0$. The second assumption implies that there exists at least one sample $s \in \bar{S}$, which does not include at least one unit of the population. We now form a sequence of integers $j_1, j_2, \cdots, j_k$ as follows. We select a sample in which at least one unit of the population is absent. If there is more than one such sample we select one of them arbitrarily. If in the selected sample more than one unit is absent, we select one of the units arbitrarily. The index of the selected unit gives j_1. Suppose $j_1, j_2, \cdots, j_m$, $m \geqq 1$ have been selcted. Let A_m, B_m be the sets (of samples) defined by,

$$A_m = \bigcup_{r=1}^{m} \bar{S}^*(j_r) , \qquad B_m = \bar{S} - A_m . \tag{4}$$

Then j_{m+1} is selected if there exists a unit U_j which is absent in at least one sample belonging to the set A_m and also in at least one sample belonging to the set B_m. If there is more than one such unit U_j then, one of the units is selected arbitrarily. The index of the selected unit gives j_{m+1}. Suppose the procedure terminates on obtaining the term j_k, i.e. it is not possible to find j_{k+1} by the procedure. Put

$$G_1 = \bigcup_{r=1}^{r=k} \bar{S}^*(j_r) , \tag{5}$$

and

$$G_2 = \bar{S} - G_1 . \tag{6}$$

Since the set $\bar{S}(j_1)$ contains at least one sample s, G_1 by (5) is nonempty. G_2 also is nonempty, because otherwise $G_1 = \bar{S}$ and comparing (5) and (3), condition C_1 is seen to be satisfied with $j_1, j_2, \cdots, j_k$ as the sequence of integers. But by assumption C_1 is not satisfied. Hence G_2 also is non-empty. Next, the units with the indices $j_1, j_2, \cdots, j_k$ are such that each of them is absent in at least one sample $s \in G_1$. There may (or may not) be some more such units. Let g_2 denote the set of all the units, such that for each unit there is at least one sample $s \in G_1$, in which that unit does not occur. Put,

$$g_1 = \mathscr{U} - g_2 . \tag{7}$$

Then by the definition of g_2, every unit in g_1, occurs in every $s \in G_1$. (The set g_1 is empty if $g_2 = \mathscr{U}$.)

Lastly all the units in g_2 must be present in every sample $s \in G_2$, because otherwise there will be one $s \in G_2$, in which some unit $U_j \in g_2$ is absent. But then the two sets shown in (4) with m replaced by k, have the sample s in common in which the unit U_j is absent, so that j would form the term j_{k+1} of the sequence $j_1, j_2, \cdots$. But by assumption the sequence terminates at j_k. Hence all the units in g_2 are present in every $s \in G_2$. Thus the partitioning defined by (5), (6) and (7) satisfies all the requirements of condition C_2.

We shall now prove the 'only if' part, and assume accordingly that the sampling design satisfies condition C_2. Suppose now that condition C_1 is satisfied in addition to C_2. The unit with the first index i_1 in the sequence in condition C_1, must belong either to the set g_1 or to g_2. Suppose it belongs to g_1 and i_m, $2 \leqq m \leqq k$,

is the first index of a unit which belongs to the set g_2. But then

$$\bigcup_{r=1}^{r=m-1} \bar{S}^*(i_r) \subset G_2 \qquad \text{by (i) of } C_2 , \tag{8}$$

and

$$\bar{S}^*(i_m) \subset G_1 \qquad \text{by (ii) of } C_2 ,$$

and since G_1 and G_2 are disjoint, the sets in the left-hand side of (8) have no sample in common which contradicts condition C_1. If alternatively none of the units with indices $i_1, i_2, \cdots, i_k$, belong to g_2, then

$$\bigcup_{r=1}^{r=k} \bar{S}^*(i_r) \subset G_2 \neq \bar{S} \tag{9}$$

which again contradicts (3) of condition C_1. A similar contradiction obviously results if the unit with the index i_1 is assumed to belong to g_2. Hence if C_2 holds C_1 cannot hold and also by the stipulation in C_2, the sampling design is not of the class $\mathscr{D}_0$. This completes the proof of Proposition 3.1.

We shall next show that if condition C_2 is satisfied, there exists an infinity of hyperadmissible estimators.

Put

$$p_1 = \sum_{s \in G_1} P_s , \qquad p_2 = \sum_{s \in G_2} P_s . \tag{10}$$

Next, taking any arbitrary real number a_1, we determine a_2 by

$$p_1 a_1 + p_2 a_2 = 0 . \tag{11}$$

This is possible, because by condition C_2, G_1, and G_2 are nonempty, so that we have in (10),

$$p_1 > 0 , \qquad p_2 > 0 .$$

For any point $\mathbf{Y} \in R_N$, let $\alpha = \alpha(s, \mathbf{Y})$ denote the set of indices of the distinct units U_i which appear in the sample s and for which the coordinate $Y_i \neq 0$. We now define an estimator T by

$$T(s, \mathbf{Y}) = \sum_{i \in \alpha} \frac{Y_i}{\pi_i} + K(\alpha) \text{ if } \alpha \text{ is nonempty} , \tag{12i}$$

$$= a_1 \text{ if } \alpha \text{ is empty and } s \in G_1 , \tag{12ii}$$

and

$$= a_2 \text{ if } \alpha \text{ is empty and } s \in G_2 . \tag{12iii}$$

In the right-hand side of (12) in the second term $K(\alpha)$ when α is nonempty depends on α only, i.e. it has the same value for all $(s, \mathbf{Y})$ which yield a given set α.

The values of $K(\alpha)$ are determined by the requirement that T is an unbiased estimate of the population total Y. By virtue of (11), (12ii) and (12iii) T is unbiased at the origin. Consider next the unbiasedness of T in the co-ordinate subspace $R(i)$, defined by $Y_i \neq 0$, $Y_j = 0$ $j \neq i$. We have

$$\pi_i K(i) + (1 - \pi_i)\, b = 0\,, \qquad \text{where} \tag{13}$$

$$b = a_2 \quad \text{by (i) of } C_2 \text{ if } U_i \in g_1$$
$$= a_1 \quad \text{by (ii) of } C_2 \text{ if } U_i \in g_2\,.$$

Hence by virtue of (1), (13) determines $K(i)$ for all i, $i = 1, 2, \cdots, N$.

Next consider a pair i, j where $\pi_{ij} > 0$. Consider the unbiasedness of T in the subspace $R(i, j)$ defined by $Y_i \neq 0$, $Y_j \neq 0$, and $Y_k = 0$ otherwise. We obtain

$$\pi_{ij} K(i, j) + (\pi_i - \pi_{ij})K(i) + (\pi_j - \pi_{ij})K(j) + (1 - \pi_i - \pi_j + \pi_{ij})\, b = 0 \tag{14}$$

where

$$b = a_2 \text{ if } U_i \in g_1\,, \qquad U_j \in g_1 \text{ by (i) of } C_2$$
$$= a_1 \text{ if } U_i \in g_2\,, \qquad U_j \in g_2 \text{ by (ii) of } C_2$$

and

$$1 - \pi_i - \pi_j + \pi_{ij} = 0\,, \qquad \text{otherwise.}$$

Note that in the last case when one of the pair of units U_i, U_j belongs to g_1 and the other to g_2 every sample $s \in \bar{S}$, contains one of the units so that the total probability of the samples which do not include either U_i or U_j is zero.

Clearly in this manner by successively increasing the size of α and considering unbiasedness of T in the subspace $R(\alpha)$, we can determine $K(\alpha)$ for all possible α, i.e. for each α for which there exists at least one $s \in \bar{S}$, such that all the units in α appear in s.

Determining $K(\alpha)$ is this manner the estimator T is unbiased. Now consider its admissibility in the subspace $R(i_1, i_2, \cdots, i_m)$ defined by $1 \leqq i_1, i_2, \cdots, i_m \leqq N$, $Y_j \neq 0$ if $j \in [i_1, i_2, \cdots, i_m]$ and $Y_j = 0$ otherwise. By (12i), at any point $\mathbf{Y}$ of this subspace, the estimator $T(s, \mathbf{Y})$ reduces to the same linear expression in Y_i for all samples $s \in \bar{S}$, for which the set α is the same and α is nonempty.

Next consider the samples s for which the set α is empty. If the units with indices $i_1, i_2, \cdots, i_m$, all belong to g_1 then by (i) of C_2 $s \in G_2$, and for all such samples the estimator has the common value a_2 by (12iii). If the units with indices $i_1, i_2, \cdots, i_m$ all belong to g_2, then by (ii) of C_2, $s \in G_1$, and again for all such s, the estimator has the common value a_1 by (12ii). Lastly if the set of units with indices $i_1, i_2, \cdots, i_m$, includes a unit belonging to g_1 and also a unit belonging to g_2, then by (i) and (ii) of C_2, there is no sample $s \in \bar{S}$ which does not include any unit belonging to this set, i.e. there is no sample for which α is empty.

Thus in all cases, for each point $\mathbf{Y}$, the value of the estimator reduces to the same linear expression in Y_i for all samples $s \in \bar{S}$, which have the same set α. Hence as shown by Hanurav ([1], Section 2) for any convex loss function, the estimator T is admissible in every subspace $R(i_1, i_2, \cdots, i_m)$ and also in R_N. It

is therefore hyperadmissible and since a_1 in (11) can have any arbitrary value, there is an infinity of hyperadmissible estimators.

Thus when the condition C_2 is satisfied, there exist infinitely many hyperadmissible estimators. When C_1 is satisfied the H-T estimator is as shown in [2] the unique hyperadmissible estimator. It is also easily seen to be the unique hyperadmissible estimator for sampling designs of the class $\mathscr{D}_0$. Hence by Proposition 3.1, the necessary and sufficient condition for the uniqueness of the H-T estimator may be stated as:

C_3: The sampling design satisfies the condition C_1 or belongs to the class $\mathscr{D}_0$; and also equivalently as

C_4: The sampling design does not satisfy the condition C_2.

The test C_2 is generally easier to apply to any given design.

A special case: As stated in condition C_2, one of the sets g_1 and g_2 may in a special case be empty. It is easily seen that such a partitioning of the population $\mathscr{U}$ exists, if and only if, $\bar{S}$ contains at least one sample which includes all the population units and also at least one sample in which all the population units do not occur. We then take G_1 to be the set of all the samples $s \in \bar{S}$, which include all the population units, G_2 the set of all $s \in \bar{S}$, which do not include all the population units, and $g_1 =$ the whole population, so that g_2 is an empty set. In all other cases, the sets g_1 and g_2 are both nonempty.

Acknowledgement. I am very grateful to the referee for pointing out some errors in the proof. I am also grateful to the referee of my previous paper [2] for suggesting this investigation.

REFERENCES

[1] HANURAV, T. V. (1968). Hyperadmissibility of estimators for finite populations. *Ann. Math. Statist.* **39** 621–642.

[2] JOSHI, V. M. (1971). Hyperadmissibility of estimators for finite populations. *Ann. Math. Statist.* **42** 680–690.