The Annals of Statistics
1976, Vol. 4, No. 5, 1012-1017

UNBIASED ESTIMATION IN FIXED COST SEQUENTIAL SAMPLING SCHEMES

BY P. K. PATHAK

The University of New Mexico

Fixed cost sequential sampling schemes are introduced in this article. In these schemes units are observed sequentially according to a given sampling method until the total cost reaches a preassigned value; it is assumed that the cost of examining each unit is unknown in advance. It is shown how the notion of sufficiency in sampling can be used to construct unbiased estimators of population parameters under these schemes.

1. Statement of the problem. Most problems of sampling from finite populations involve the determination of the optimum sample size. In problems where estimators with prescribed precision are required, the sample size is usually determined so as to obtain estimators with maximum precision at a fixed cost. In many of these problems the determination of the sample size is often based on the assumption that the cost of observing a unit in the population is nearly the same for all population units; and sampling is planned so as to reach some expected cost for the survey. In populations where the cost of observing units varies greatly from unit to unit, sampling procedures of the above kind lead to highly variable random costs of sample selection, a feature which is perhaps not very desirable in fixed cost surveys at least. For fixed cost surveys it would, therefore, be desirable to have sequential sampling procedures which eliminate the randomness of the total cost of sample selection. To this end we introduce fixed cost sampling schemes in this paper. In these sampling schemes units are observed sequentially according to a given sampling method and sampling stopped when the accumulated cost reaches the preassigned total cost; it is assumed that the cost of examining each unit is unknown in advance. The primary object of this article is to show how the notion of sufficiency can be used to construct efficient unbiased estimators of population parameters under these schemes. It turns out that customary estimators of population parameters under nonsequential sampling schemes continue to remain efficient estimators under the corresponding fixed cost sequential sampling schemes. For reasons of simplicity we shall illustrate the technique of construction of unbiased estimators in the case of fixed cost simple random sampling only.

2. Fixed cost simple random sampling. Consider a population of N elements. Suppose that the jth population unit is $U_j = (j, Y_j)$, where j is its unit-index and $Y_j = (C_j, Z_j)$ where Z_j is a real-valued variate of interest and C_j is the cost

Received April 8, 1965; revised January 12, 1976.

AMS 1970 *subject classifications*. Primary 62B05, 62D05; Secondary 62L12.

Key words and phrases. Fixed cost sampling, sufficiency in sampling, statistic of distinct units, minimal sufficient statistic, unbiased estimation, inverse sampling.

(possibly unknown) of ascertaining the value of Z_j $(1 \leqq j \leqq N)$. (It is perhaps worthwhile mentioning here that it takes a certain amount of time and effort to examine the U_j completely and thus ascertain the value of Z_j; this fact is tacitly incorporated in the cost of selection C_j. Before sampling one does not normally know C_j but once the U_j has been examined completely, one knows the true value of C_j.) We consider here the problem of estimating the population mean

$$\bar{Z} = N^{-1} \sum_{j=1}^{N} Z_j \tag{2.1}$$

and the population variance

$$S^2 = (N-1)^{-1} \sum_{j=1}^{N} (Z_j - \bar{Z})^2 . \tag{2.2}$$

It is assumed that the total cost, say L, to be spent on sample selection is fixed in advance. We also assume $C_j + C_k < L$ for all j and k with $j \neq k$, $C_1 + C_2 + \cdots + C_N > L$, $\min(C_1, \cdots, C_N) > 0$, and the C_j are otherwise unknown. (The case in which costs $C_1, \cdots, C_N$ are known can be handled by a nonsequential scheme.)

The following sampling procedure is to be adopted. Population units are drawn one by one with equal probabilities without replacement until the accumulated cost reaches L. More precisely, the sample is $(X_1, \cdots, X_M)$ where each X_j is one of the U_j and the stopping variable M is defined as follows: $M = r$ if and only if $\sum_1^{(r-1)} C(X_i) < L$ and $\sum_1^r C(X_i) \geqq L$, $C(X_i)$ being the cost of observing the ith unit completely. (Note that the X_j are distinct because sampling is without replacement.)

Now from the viewpoint of applications it is desirable to treat X_M differently from $X_1, \cdots, X_{M-1}$ since because of cost overrun it may be decided not to examine X_M (i.e., ascertain its z-value) completely. Also a second reason, as we shall see later, is that treating X_M differently leads to substantial mathematical simplicity in the construction of unbiased estimators. Consequently we define

$$T_M = (M, \{X_1, \cdots, X_{M-1}\}, X_M) . \tag{2.3}$$

Also we denote the observed sample by

$$S_M = (X_1, \cdots, X_M) \tag{2.4}$$

and the statistic of distinct units by

$$T_0 = \{X_1, \cdots, X_M\} . \tag{2.5}$$

(Notice that S_M is a sequence of M units while T_0 is the set whose elements are $X_1, \cdots, X_M$.)

We briefly turn now to the notion of sufficiency in sampling which will be the basis of our study in the sequel. The *statistic of distinct units* T_0 plays a central role in the study of sufficiency in sampling. It is rather easy to see, using conditional probability distributions, that T_0 is a sufficient statistic. In a number of special cases the sufficiency of T_0 was first noted by Basu [1] and Hájek [4]. In a general setting that includes sampling under arbitrary (generally sequential)

sampling schemes as a special case, Basu and Ghosh [2] and Basu [3] have now established the following results concerning sufficiency in sampling:

1. The notions of sufficient statistic, sufficient partitions and sufficient subfields are all equivalent in the general framework of sample surveys.

2. Under the added assumption that for each j, $1 \leqq j \leqq N$, Y_j can take on at least two different values for each set of possible values of the other Y's, the statistic T_0 *is minimal sufficient*, and a statistic T is sufficient if and only if for all samples s_1, s_2 the equality $T(s_1) = T(s_2)$ implies $T_0(s_1) = T_0(s_2)$.

In 1965 partial results along these lines were first obtained by the writer and submitted to the *Annals of Mathematical Statistics*. Specifically it was shown that in nonsequential sampling schemes a partition of the sample space into blocks (atoms) is sufficient if and only if the statistic T_0 remains constant over every block of the partition, i.e., if samples s_1 and s_2 belong to the same block then $T_0(s_1) = T_0(s_2)$. As a consequence of this characterization of sufficient partitions, it was noted that under a very mild condition on the parameter space, the statistic T_0 of distinct units together with their Y-values induces the minimal sufficient partition. Publication of these results was delayed, and the results have now been superseded by the more general and elegant results of [2] and [3].

We turn now to applications of the above-mentioned characterization of sufficiency to unbiased estimation in fixed cost simple random sampling. We note that since the minimal sufficient statistic T_0 given by (2.5) is a function of the statistic T_M given by (2.3), it follows from the preceding discussion that T_M is also sufficient. In Theorem 2.1 below we derive an unbiased estimator of $\bar{Z} = N^{-1} \sum Z_j$ by starting from the unbiased estimator $Z(X_1)$ based on the first sample unit X_1 and taking its conditional expectation given T_M (Rao–Blackwellization). An estimator for the variance of the estimator for $\bar{Z}$ is obtained in a similar way.

THEOREM 2.1. *In fixed cost simple random sampling, an unbiased estimator of the population mean* $\bar{Z}$ *is given by*

$$\bar{Z}_{(M-1)} = (M-1)^{-1} \textstyle\sum_1^{(M-1)} Z(X_i) \tag{2.6}$$

where $Z(X_i)$ *denotes the Z-characteristic of the* ith *sample unit. Further the variance of* $\bar{Z}_{(M-1)}$ *is given by*

$$V(\bar{Z}_{(M-1)}) = \frac{1}{2N(N-1)} \textstyle\sum_{j \neq j'=1}^{N} (Z_j - Z_{j'})^2 \times \left[E\left[\frac{1}{M-1} \,\middle|\, X_1 = U_j, X_2 = U_{j'} \right] - \frac{1}{N} \right], \tag{2.7}$$

and an unbiased estimator of $V(\bar{Z}_{(M-1)})$ *is given by*

$$v(\bar{Z}_{(M-1)}) = \left(\frac{1}{(M-1)} - \frac{1}{N} \right) \frac{1}{(M-2)} \left[\textstyle\sum_{i=1}^{(M-1)} (Z(X_i) - \bar{Z}_{(M-1)})^2 \right]. \tag{2.8}$$

PROOF. Clearly $Z(X_1)$ is an unbiased estimator of Z. We proceed to compute its conditional expectation given T_M. Let $k > 1$, let $U_{i_1}, \cdots, U_{i_k}$, U_j be $(k+1)$

distinct population units and suppose that

$$T_M = (k+1, \{U_{i_1}, \cdots, U_{i_k}\}, U_j)$$

is given. Then the conditional distribution of X_1 is easily seen to be concentrated on the set $\{U_{i_1}, \cdots, U_{i_k}\}$. By the symmetry of the cost function in $\{U_{i_1}, \cdots, U_{i_k}\}$ it follows that

$$P[X_1 = U_{i_1} \mid T_M] = \cdots = P[X = U_{i_k} \mid T_M] = \frac{1}{k}.$$

So

$$\begin{aligned} E[Z(X_1) \mid T_M = (k+1, \{U_{i_1}, \cdots, U_{i_k}\}, U_j)] \\ = \frac{1}{k} \textstyle\sum_1^k Z(U_{i_1}) = (M-1)^{-1} \sum_1^{(M-1)} Z(X_i) \end{aligned}$$

is an unbiased estimator of $\bar{Z}$.

We now derive (2.8). We note that if $e(\bar{Z}^2)$ is an unbiased estimator of $\bar{Z}^2$ then

$$v(\bar{Z}_{(M-1)}) = \bar{Z}^2_{(M-1)} - e(\bar{Z}^2) \tag{2.9}$$

is an unbiased estimator or $V(\bar{Z}_{(M-1)})$. Since

$$\begin{aligned} \bar{Z}^2 &= N^{-2} \textstyle\sum_{j=1}^N Z_j^2 + N^{-2} (\sum_{j \neq j'=1}^N Z_j Z_{j'}), \\ t_2 &= N^{-1} Z^2(X_1) + (N-1) N^{-1} Z(X_1) Z(X_2) \end{aligned}$$

is an unbiased estimator of $\bar{Z}^2$ based on the first two units. It therefore follows that $E[t_2 \mid T_M]$ would be a reasonable unbiased estimator of $\bar{Z}^2$. It is easily seen that

$$\begin{aligned} &E[N^{-1} Z(X_1) + N^{-1}(N-1) Z(X_1) Z(X_2) \mid T_M] \\ &= \frac{1}{N(M-1)} \textstyle\sum_1^{(M-1)} Z(X_i)^2 \\ &\quad + \frac{(N-1)}{N(M-1)(M-2)} \textstyle\sum_{i \neq i'=1}^{(M-1)} Z(X_i) Z(X_{i'}) \\ &= \bar{Z}^2_{(M-1)} - \left[\frac{1}{(M-1)} - \frac{1}{N}\right] \frac{1}{(M-2)} \textstyle\sum_{i=1}^{(M-1)} (Z(X_i) - \bar{Z}_{(M-1)})^2 . \end{aligned} \tag{2.10}$$

Substituting this expression for $e(\bar{Z}^2)$ into (2.9) yields (2.8).

We now use (2.8) to deduce (2.7) as follows:

$$\begin{aligned} V(\bar{Z}_{(M-1)}) &= E[v(\bar{Z}_{(M-1)})] \\ &= E\left[E\left[\left(\frac{1}{(M-1)} - \frac{1}{N}\right) \right.\right. \\ &\qquad \left.\left. \times \frac{1}{2(M-1)(M-2)} \textstyle\sum_{i \neq i'=1}^{(M-1)} (Z(X_i) - Z(X_{i'}))^2 \mid T_M\right]\right]. \end{aligned} \tag{2.11}$$

Since $X_1, \cdots, X_{(M-1)}$ are interchangeable for a given M, (2.11) becomes

$$\begin{aligned} V(\bar{Z}_{(M-1)}) &= E\left[E\left[\left(\frac{1}{(M-1)} - \frac{1}{N}\right) \tfrac{1}{2}(Z(X_1) - Z(X_2)^2 \,\middle|\, T_M\right]\right] \\ &= E\left[\left(\frac{1}{(M-1)} - \frac{1}{N}\right) \tfrac{1}{2}(Z(X_1) - Z(X_2))^2\right]. \end{aligned}$$

Thus

$$V(\bar{Z}_{(M-1)}) = E\left[\tfrac{1}{2}\textstyle\sum_{j\neq j'=1}^{N}(Z_j - Z_{j'})^2\left(\frac{1}{(M-1)} - \frac{1}{N}\right)\alpha_{jj'}\right] \tag{2.12}$$

where $\alpha_{jj'} = 1$ if $X_1 = U_j$ and $X_2 = U_{j'}$, and $= 0$ otherwise. It is easily seen that

$$E[\alpha_{jj'}] = P[X_1 = U_j,\ X_2 = U_{j'}] = \frac{1}{N(N-1)} \tag{2.13}$$

and

$$\begin{aligned} E\left[\frac{1}{(M-1)}\alpha_{jj'}\right] &= E\left[E\left[\frac{1}{(M-1)}\alpha_{jj'}\,\middle|\, X_1, X_2\right]\right] \\ &= \frac{1}{N(N-1)}\cdot E\left[\frac{1}{(M-1)}\,\middle|\, X_1 = U_j,\ X_2 = U_{j'}\right]. \end{aligned} \tag{2.14}$$

Substituting (2.13) and (2.14) in (2.12), we obtain (2.7). □

REMARKS. (i) In a similar fashion it can be shown that

$$s^2_{(M-1)} = \frac{1}{(M-2)}\textstyle\sum_{i=1}^{(M-1)}(Z(X_i) - \bar{Z}_{(M-1)})^2$$

is an unbiased esitimator of the population variance S^2 give by (2.2).

(ii) The estimator $\bar{Z}_{(M-1)}$ considered here is inadmissible since it ignores the last sample unit X_M. Nonetheless the loss in efficiency should be small if M is large. An estimator better than $\bar{Z}_{(M-1)}$ would be $E[Z(X_1)\,|\,\{X_1, \cdots, X_M\}]$. This latter estimator does not have a simple expression, and is, consequently, of little use in practice. It is, however, possible to construct a simple estimator better than $\bar{Z}_{(M-1)}$ by conditioning $Z(X_1)$ with respect to the following statistic:

$$\begin{aligned} T^* &= T_M && \text{if} \quad C(X_1) + \cdots + C(X_M) > L\,, \\ &= \{X_1, \cdots, X_M\} && \text{if} \quad C(X_1) + \cdots + C(X_M) = L\,. \end{aligned}$$

Since T_M and $\{X_1, \cdots, X_M\}$ are both sufficient, it follows from Theorem 5 ([5], page 323) that T^* is a sufficient statistic. It can be shown that

$$\begin{aligned} E[Z(X_1)\,|\,T^*] &= (M-1)^{-1}\textstyle\sum_{i=1}^{(M-1)} Z(X_i) && \text{if} \quad C(X_1) + \cdots + C(X_M) > L\,, \\ &= M^{-1}\textstyle\sum_{i=1}^{M} Z(X_i) && \text{if} \quad C(X_1) + \cdots + C(X_M) = L\,. \end{aligned}$$

Since T^* is a function of T_M,

$$E[Z(X_1)\,|\,T^*] = E[E[Z(X_1)\,|\,T_M]\,|\,T^*] = E[\bar{Z}_{(M-1)}\,|\,T^*]\,.$$

So $E[Z(X_1)\,|\,T^*]$ is a better estimator than $\bar{Z}_{(M-1)}$.

(iii) In a similar manner one can consider fixed cost sampling schemes in situations where population units are selected by other sampling methods such as sampling with unequal probabilities (with replacement), and multi-stage sampling schemes, etc. The estimators of population parameters under their fixed cost analogues will be similar to their corresponding estimators in fixed sample size procedures. It is also possible to introduce fixed cost sampling when

sampling is carried out from arbitrary probability distributions. We omit the details for reasons of brevity.

(iv) It is remarked that the notion of cost of selection of units introduced in this paper is quite general and does not necessarily have anything to do with the actual cost of selection. For instance cost can be replaced by the amount of time taken to select units or any other nonnegative index associated with the units. Techniques analogous to those considered in this section can be used to construct unbiased estimators of population parameter in these situations. For illustration it is shown below that similar techniques apply to inverse sampling with unequal probabilities.

Under inverse sampling with unequal probabilities, units are drawn sequentially one-by-one with unequal probabilities (with replacement). The stopping variable M is such that the sampling is stopped at $M = (k + 1)$ if X_{k+1} is the first $(n + 1)$st distinct unit, where n is a given positive integer. In this scheme an unbiased estimator of the population total $Z = \sum_{j=1}^{N} Z_j$ is given by

$$Z_{(M-1)} = \frac{1}{(M-1)} \sum_{i=1}^{(M-1)} Z(X_i)/P(X_i)$$

where $P(X_i)$ is the probability of selection of X_i. The derivation of this estimator is analogous to that of the corresponding estimator $Z_{(M-1)}$ of fixed cost simple random sampling. For further details on inverse sampling the reader may refer to the papers [6] and [7].

3. Acknowledgment. The author is grateful to Professor R. A. Wijsman for his help in revising the paper.

REFERENCES

[1] BASU, D. (1958). On sampling with and without replacement. *Sankhyā* **20** 287–294.

[2] BASU, D. and GHOSH, J. K. (1967). Sufficient statistics in sampling from a finite universe. *Bull. Inst. Internat. Statist.* **42** 850–859.

[3] BASU, D. (1969). Role of sufficiency and likelihood principle in sample survey theory. *Sankhyā A* **31** 441–454.

[4] HÁJEK, JAROSLAV (1959). Optimum strategy and other problems in probability sampling. *Časopis Pěst. Mat.* **84** 387–423.

[5] PATHAK, P. K. (1962). On sampling with unequal probabilities. *Sankhyā A* **24** 315–326.

[6] PATHAK, P. K. (1964). On inverse sampling with unequal probabilities. *Biometrika* **51** 185–193.

[7] SAMPFORD, M. R. (1962). Methods of cluster sampling with and without replacement for clusters of unequal sizes. *Biometrika* **49** 27–40.

THE UNIVERSITY OF NEW MEXICO
ALBUQUERQUE, NEW MEXICO 87131