The Annals of Statistics
1976, Vol. 4, No. 5, 993–997

ON UNIFORMLY MINIMUM VARIANCE ESTIMATION IN FINITE POPULATIONS

BY CARL ERIK SÄRNDAL

University of British Columbia

In the literature one finds (at least) two approaches towards proving that the sample mean is uniformly minimum variance (UMV), among unbiased estimates that "ignore the labels," for the finite population mean: The "traditional approach" and the "scale-load approach." The identity of results under the two approaches extends to a more general setting, as shown in this paper: The Horvitz-Thompson estimate is UMV unbiased for any given fixed effective size design.

1. Introduction. Consider a population of N units, each unit being identified by a label k; $k = 1, \cdots, N$. Let $\mathscr{S}$ be the set of subsets $\{s_n\}$, where each s_n contains n, a fixed number, of labels drawn from the set $\{1, \cdots, N\}$. Let $p(s_n)$ be a given function on $\mathscr{S}$ such that $\sum_{s_n \in \mathscr{S}} p(s_n) = 1$. We are assuming (throughout the paper) a *fixed effective* size design $p(\cdot)$, i.e., $p(\cdot)$ assigns nonzero probability only to sets containing exactly n distinct labels.

Considering that "mass-draw" of n units may not be practical, assume that the given design $p(\cdot)$ is implemented through a without replacement, draw-by-draw mechanism now to be described:

For $i = 1, 2, \cdots, n-1$, let s_i denote any set of i distinct labels, and let $p_i(s_i) = \sum p(s_n)/\binom{n}{i}$, where $\sum$ is over those $\binom{N-i}{n-i}$ sets s_n of which s_i is a subset; for some of the sets s_n, $p(s_n) = 0$ is, of course, a possibility. For each i, $\sum p_i(s_i) = 1$, where $\sum$ is over the $\binom{N}{i}$ different subsets s_i. In particular, if $i = 1$ and k is the only element of s_1, we have $p_1(s_1) = \alpha_k/n$, where α_k denotes the inclusion probability of label k.

The draw-by-draw mechanism, to be denoted M, is defined as follows:

First draw label k_1 with probability $p_1(s_1) = \alpha_{k_1}/n$, where $s_1 = \{k_1\}$. Then, in the ith draw, $i = 2, 3, \cdots, n$, and given that $s_{i-1} = \{k_1, \cdots, k_{i-1}\}$ resulted from the first $i-1$ draws, draw label k_i, for $k_i = 1, \cdots, N$, $k_i \notin s_{i-1}$, with probability $p_i(s_i)/ip_{i-1}(s_{i-1})$, where $s_i = \{k_1, \cdots, k_{i-1}, k_i\}$ and $p_n(\cdot) = p(\cdot)$. Clearly, the probability of obtaining the labels of a given set s_n in any one particular of the $n!$ drawing orders will be $p(s_n)/n!$.

From here on, we write simply s (in place of s_n) for the set of labels resulting from the draw-by-draw sampling, and $\sum_s$ for summation over $k \in s$.

Assume that y, the character of interest, takes the value y_k for unit k. We shall consider the conditions under which the Horvitz–Thompson (HT) estimate

Received May 1975; revised January 1976.

AMS 1970 *subject classifications.* Primary 62005; Secondary 62F10.

Key words and phrases. Estimation, uniformly minimum variance, Horvitz–Thompson estimator, scale-loads, labels, sufficiency, completeness.

$t_{HT} = N^{-1} \sum_s y_k/\alpha_k$, or its special case, $\bar{y}_s = n^{-1} \sum_s y_k$, has the uniformly minimum variance (UMV) property in unbiased estimation of $\bar{Y} = N^{-1} \sum_{k=1}^{N} y_k$.

We may distinguish two approaches towards proving the UMV property. In the *traditional approach*, the stochastic element enters by considering the measurement y_k, or some function thereof, as the outcome of a random variable η_i, $i = 1, \cdots, n$. In the *scale-load approach* of Hartley and Rao (1968), Royall (1968), $y_{01}, \cdots, y_{0J}$ ($J \leqq N$), say, denote the distinct numbers among $y_1, \cdots, y_N$, and the vector of sample frequencies of $y_{01}, \cdots, y_{0J}$, $\mathbf{n}_s = (n_1, \cdots, n_J)$, is treated as the outcome of the discrete random vector, $\boldsymbol{\nu}_s = (\nu_1, \cdots, \nu_J)$.

Neyman (1934) showed, in the spirit of the traditional approach, that the sample mean $\bar{y}_s$ is UMV in the class of unbiased linear estimates, $\sum_{i=1}^{n} c_i y_{k_i}$, under simple random sampling without replacement (srs). Removing the restriction to linearity, Watson (1964) showed this result to hold, under srs, in the class of arbitrary unbiased functions of y_{k_i}, $i = 1, \cdots, n$. Hartley and Rao (1968) used the scale-load approach to show, under srs, the UMV property of $\bar{y}_s$ in the class of arbitrary unbiased functions of the scale-point frequencies $n_1, \cdots, n_J$. While their result is identical to the unpublished result of Watson (1964), their approach may be seen as an *alternative* means of justifying the sample mean.

Later, Hartley and Rao (1969) considered the distinct numbers among $r_k = y_k/a_k$, $k = 1, \cdots, N$, as new scale points, under sampling with replacement (hence not a fixed effective size design), a_k being the probability of drawing unit k, in each of n independent draws. They showed, among other things, that the sample frequencies $n_1, \cdots, n_J$ are sufficient for the corresponding unknown population frequencies of the new scale points.

The traditional approach was used in Särndal (1972) to show that t_{HT} is UMV in the unbiased class linear in y_{k_i}/α_{k_i}, $i = 1, \cdots, n$.

The main result shown below (Theorem 1) is that the scale-load approach of Hartley and Rao (1969) can be carried through, for a fixed effective size design, to show the UMV property of the HT estimate. We also note that the same conclusion obtains through the traditional approach, i.e., the restriction to linearity in the result of Särndal (1972) can be removed.

These results do not contradict the well-known fact that in a more general, label dependent class no UMV unbiased estimate exists (Godambe (1955)). Any UMV result established within a less general class, including all the results mentioned earlier in this paper, is therefore in a sense limited, cf. the discussion in Godambe (1970). In spite of such limitations, it is interesting that both scale-load and traditional approaches do admit the interpretation of the HT estimate as being UMV, in the sense specified in this paper. Considering the recent strong interest in the foundations of survey sampling, it is important to lay down the exact conditions under which UMV unbiased estimation for finite populations is indeed possible.

2. UMV estimation in the scale-load approach. Consider the *scale-load approach.* Assume among the numbers $z_k = ny_k/N\alpha_k$, $k = 1, \cdots, N$, there are $J \leqq N$ distinct ones, say, $b_1, \cdots, b_J$. For $j = 1, \cdots, J$, set $u_j = \{k : z_k = b_j\}$, $N_j =$ number of elements k in u_j, and $A_j = \sum_{k \in u_j} \alpha_k/n$. Hence, $\sum_{j=1}^J N_j = N$; $A_j \geqq 0$, $j = 1, \cdots, J$, and $\sum_{j=1}^J A_j = 1$. The task is to estimate $\bar{Y} = \sum_{j=1}^J A_j b_j$, where the b_j are *fixed* numbers and the A_j are *unknown parameters.*

For $s \in \mathscr{S}$ and $j = 1, \cdots, J$, set $n_j =$ number of $k \in s$ such that $z_k = b_j$; hence $\sum_{j=1}^J n_j = n$. We prove the following:

THEOREM 1. *For any given fixed effective size design $p(s)$, implemented by the mechanism M and such that $\alpha_k > 0$, $k = 1, \cdots, N$, the* HT *estimate*

$$t_{\mathrm{HT}} = N^{-1} \sum_s y_k/\alpha_k = n^{-1} \sum_{j=1}^J n_j b_j ,$$

is UMV *for $\bar{Y}$ in the class of unbiased estimates consisting of arbitrary functions of $n_1, \cdots, n_J$.*

The proof is accomplished by showing three things: (a) that n_j/n is unbiased for A_j, $j = 1, \cdots, J$, whereby it will follow that $\sum_{j=1}^J n_j b_j/n$ is unbiased for $\sum_{j=1}^J A_j b_j$; (b) that $\mathbf{n}_s = (n_1, \cdots, n_J)$ is a sufficient statistic for $A_1, \cdots, A_J$; and (c) that $\mathbf{n}_s$ is complete.

Unbiasedness. Consider u_j, containing N_j labels. Let $m = \min(n, N_j)$. For $n_{0j} = 0, 1, \cdots, m$, set $\mathscr{S}_{0j} = \{s : n_j = n_{0j}\}$ and $q(n_{0j}) = \sum_{s \in \mathscr{S}_{0j}} p(s)$. The unbiasedness follows from

$$n^{-1}E(n_j) = n^{-1} \sum_{v=0}^m vq(v) = n^{-1} \sum_{k \in u_j} \sum_{s \supset k} p(s) = n^{-1} \sum_{k \in u_j} \alpha_k = A_j .$$

Sufficiency of $\mathbf{n}_s$. The observable result of the sampling is a sequence of n pairs, (k_i, b_{j_i}), where $b_{j_i} = z_{k_i}$, $i = 1, \cdots, n$. If we ignore the label part, the sequence $\mathbf{b}_s$, consisting of the n numbers b_{j_i}, remains. Since each sequence has the same probability, we get $p(\mathbf{b}_s \mid \mathbf{n}_s) = \prod_{j=1}^J n_j!/n!$. This does not depend on the A_j, hence the suffiency, i.e., the drawing order is unimportant.

Completeness. First, let $J = 2$, i.e., $z_k = b_1$ for N_1 labels, and $z_k = b_2$ for the rest. The possible values of nA_1 are 0 (if $N_1 = 0$); α_k, $k = 1, \cdots, N$ (if $N_1 = 1$ and $z_k = b_1$); $\alpha_k + \alpha_l$, $k \neq l = 1, \cdots, N$ (if $N_1 = 2$ and $z_k = z_l = b_1$), etc. We must show that $E[g(n_1)] = 0$ for a real function g and all possible values of A_1 implies $g(n_1) = 0$ for $n_1 = 0, 1, \cdots, n$. Assume without loss of generality that $p(s) > 0$ for $s = \{1, 2, \cdots, n\}$. First, consider $N_1 = 0$, i.e., $q(0) = 1$. Thus $E[g(n_1)] = 0$ implies $g(0) = 0$. Next, consider $N_1 = 1$ and $z_1 = b_1$. Then $q(1) = 1 - q(0) = \alpha_1 > 0$, and $E[g(n_1)] = 0$ implies $g(1) = 0$. The sets $\{1, 2\}, \cdots, \{1, 2, \cdots, n\}$ have nonzero probability, hence we conclude that $g(n_1) = 0$ for $n_1 = 2, \cdots, n$. To show the completeness for an arbitrary $J > 2$, a proof by induction (similar to that of S. K. Kale, cited by Hartley and Rao (1968), page 549) may be used; the details are omitted.

3. Concluding remarks.

REMARK 1. In the traditional approach, we have a result equivalent to that

of Theorem 1: For any given fixed effective size design $p(s)$, implemented by the mechanism M and such that $\alpha_k > 0$, $k = 1, \cdots, N$, the HT estimate,

$$t_{\mathrm{HT}} = N^{-1} \sum_s y_k/\alpha_k = n^{-1} \sum_s z_k ,$$

is UMV for $\bar{Y}$ in the class of unbiased estimates consisting of arbitrary functions of $z_{k_1}, \cdots, z_{k_n}$. This statement, of which the results of Neyman (1934), Watson (1964), Särndal (1972) are special cases, follows easily, letting, for $i = 1, \cdots, n$, ζ_i be the random variable that takes the value z_{k_i} if label k_i occurs in the ith draw. The probability of $\zeta_i = z_{k_i}$, $i = 1, \cdots, n$, remains $p(s)/n!$ under any permutation of $k_1, \cdots, k_n$. If $f(z_{k_1}, \cdots, z_{k_n})$ is an unbiased estimate of $\bar{Y} = \sum_{k=1}^{N} z_k \alpha_k/n$, then the symmetrized function $\sum f(z_{r_1}, \cdots, z_{r_n})/n!$, where $\sum$ is over all permutations $r_1, \cdots, r_n$ of $k_1, \cdots, k_n$, is also unbiased and has smaller variance than f, unless f is already symmetric. Thus, information about the drawing order can be discarded with no increase in variance. The remaining set of numbers, $\{z_k : k \in s\}$, is complete; this follows as an extension of Royall's (1968) completeness result.

REMARK 2. Assume now that the given fixed effective size design is implemented through "mass-draw" of the n units, i.e., by selecting the set s with probability $p(s)$. Now, in the absence of drawing order, the result of the sampling, after labels have been ignored, is the set of numbers, $\{z_k : k \in s\}$, or, looking at it from the scale-load point of view, the frequencies n_j of the scale-points b_j. The completeness results discussed above, in the scale-load approach and in the traditional approach, ensure the uniqueness of the HT estimate as an unbiased estimate of $\bar{Y}$.

REMARK 3. Consideration of the scale loads z_k obviously makes good sense only when the y_k/α_k are approximately constant (see, e.g., J. N. K. Rao (1975); a similar requirement is inherent in C. R. Rao's (1971) consideration of the HT estimate under random permutation models). As is well known, the HT estimate can be very poor if y_k and α_k are weakly or negatively correlated.

REFERENCES

[1] GODAMBE, V. P. (1955). A unified theory of sampling from finite populations. *J. Roy. Statist. Soc. Ser. B* **17** 268–278.

[2] GODAMBE, V. P. (1970). Foundations of survey sampling. *Amer. Statist.* **24** (no. 1) 33–38.

[3] HARTLEY, H. O. and RAO, J. N. K. (1968). A new estimation theory for sample surveys. *Biometrika* **55** 547–557.

[4] HARTLEY, H. O. and RAO, J. N. K. (1969). A new estimation theory for sample surveys, II. In *New Developments in Survey Sampling* (N. L. Johnson and H. Smith, Jr., eds.) 147–169. Wiley, New York.

[5] NEYMAN, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *J. Roy. Statist. Soc.* **97** 558–606.

[6] RAO, C. R. (1971). Some aspects of statistical inference in problems of sampling from finite populations. In *Foundations of Statistical Inference* (V. S. Godambe and D. A. Sprott, eds.) 177–202. Holt, Rinehart and Winston, Toronto.

[7] RAO, J. N. K. (1975). On the foundations of survey sampling. In *A Survey of Statistical Design and Linear Models* (J. N. Srivastava, ed.) 489–505. North-Holland, Amsterdam.
[8] ROYALL, R. (1968). An old approach to finite population sampling theory. *J. Amer. Statist. Assoc.* **63** 1269–1279.
[9] SÄRNDAL, C. E. (1972). Sample survey theory vs. general statistical theory: Estimation of the population mean. *Internat. Statist. Rev.* **40** 1–12.
[10] WATSON, G. S. (1964). Estimation in finite populations. (Unpublished report.)

UNIVERSITY OF BRITISH COLUMBIA
2075 WESBROOK PLACE
VANCOUVER, BRITISH COLUMBIA V6T 1W5