The Annals of Statistics
1973, Vol. 1, No. 6, 1212–1221

ESTIMATION IN SAMPLING THEORY WITH EXCHANGEABLE PRIOR DISTRIBUTIONS

BY V. P. GODAMBE AND M. E. THOMPSON

University of Waterloo

Godambe (1955) and Godambe and Joshi (1965) established the joint optimality of the Horvitz–Thompson estimates and the corresponding sampling designs with respect to a class of prior distributions which were "product measures." In the present paper the optimality is established replacing the "product measures" by prior distributions which are exchangeable in appropriately transformed variates. Furthermore, the optimality criterion is justified in terms of Chebychev's inequality.

1. Introduction and notation. In relation to estimating the total of a finite population, a *criterion* for optimality jointly of 'an estimator and a sampling design' was proposed by Godambe (1955). Thus 'an estimator and a sampling design' are said jointly to be optimal if they provide the minimal expected variance for a given *class* of prior distributions, for all sampling designs of a fixed sample size and corresponding unbiased estimators of the population total. According to this criterion, it was proved (Godambe (1955), Godambe and Joshi (1965)), that the Horvitz–Thompson estimator and a sampling design with corresponding inclusion probabilities were jointly optimum for a certain *class* of prior distributions consisting of product measures on R_N (Euclidean N-space), N being the size of the finite population. The optimality corresponding to the class of all exchangeable (exchangeable in appropriately transformed variates) prior distributions is established in this paper. Some partial results in this direction are previously due to Kempthorne (1969), Rao (1971) and Thompson (1971). A further discussion of the optimality criterion referred to above is also included in this paper.

Following the notation in the literature on the subject we denote the finite population of size N, i.e. one consisting of N individuals i, $i = 1, \cdots, N$, by $\mathscr{P}_N$. Let S denote all the subsets s of $\mathscr{P}_N$, so that $S = \{s, s \subset \mathscr{P}_N\}$.

DEFINITION 1.1. Any $s \in S$ is called a sample.

DEFINITION 1.2. Any real function p on S such that $1 \geqq p(s) \geqq 0$ for all $s \in S$ and $\sum_S p(s) = 1$ is called a *sampling design*.

DEFINITION 1.3. The real function ν on S, where $\nu(s)$ = total number of individuals i ($i = 1, \cdots, N$) such that $i \in s$ is called the *sample size*.

DEFINITION 1.4. A sampling design p (Definition 1.2) is said to be a *fixed sample size* design if for some fixed number n the probability $p(s) = 0$ whenever $\nu(s) \neq n$ for $s \in S$.

Received December 1971; revised February 1973.

Now let S_i be the subset of S, $S_i \subset S$, such that $[s \in S_i] \Leftrightarrow [s \ni i]$. That is, all samples s in S_i include the given individual i and if any s in S includes i then s is in S_i. Then we have

DEFINITION 1.5. For any sampling design p (Definition 1.2) the number $a_i = \sum_{S_i} p(s)$ is called the *inclusion probability* for the individual i ($i = 1, \cdots, N$). The inclusion probabilities a_i in Definition 1.5 satisfy for any fixed sample size design (Definition 1.4) p such that $[\nu(s) \neq n] \rightarrow [p(s) = 0]$ the following well-known relationship:

$$\sum_1^N a_i = n , \tag{1}$$

(Godambe, 1955).

Now the problem of estimation is as follows: With every individual i ($i = 1, \cdots, N$) in the population is associated a real variate x_i ($i = 1, \cdots, N$).

DEFINITION 1.6. The vector $\mathbf{x} = (x_1, \cdots, x_N)$ is called the *population vector.* We suppose that before any sampling is done (for the present occasion) this population vector $\mathbf{x}$ is unknown except for certain types of prior knowledge referred to later. With the considerations of this prior knowledge and the cost of the survey, the statistician employs a sampling design p (Definition 1.2) and draws a sample s. The values $x_i : i \in s$ are then observed by him. The problem we consider here is of estimating (on the basis of p, $(s, x_i : i \in s)$, and the prior knowledge) a function of population vector (Definition 1.6) called the population total defined by

$$T(\mathbf{x}) = \sum_1^N x_i , \qquad \mathbf{x} \in R_N . \tag{2}$$

DEFINITION 1.7. An estimator (for $T(\mathbf{x})$) is a real-valued function $e(s, \mathbf{x})$ on $S \times R_N$ which depends on $\mathbf{x}$ only through $x_i : i \in s$. We shall use the more precise notation $e(s, x_i : i \in s)$.

To see how the foregoing framework is general enough to study most of the estimation problems arising in relation to sampling finite populations, one may refer to Godambe (1965, 1969).

2. Minimum expected variance. Due to the general non-existence of uniformly minimum variance unbiased (UMVU) estimates in survey-sampling (Godambe (1955)), most often in the literature unbiased estimates with minimum expected variance are investigated. The intuitive appeal of this procedure in general can be further strengthened by recourse to the usual Chebychev's inequality: if $E(\cdot \mid \theta)$ denotes expectation given a parameter θ, for any estimator t and for all $\propto$, θ and k,

$$P(|t - \propto| \leqq k \mid \theta) \geqq 1 - \frac{E((t - \propto)^2 \mid \theta)}{k^2} . \tag{3}$$

This inequality suggests that the smaller the value of $E((t - \propto)^2 \mid \theta)$, the higher will be the probability concentration of t at $\propto$, given θ. Thus, for example, requirement that the probability concentration of t should be no less at the true value $h(\theta)$ of a certain parametric function h than at the false value $h(\theta')$ is approximately the requirement that $E((t - h(\theta))^2 \mid \theta) \leqq E((t - h(\theta'))^2 \mid \theta)$ for all θ and θ'. This, under rather trivial assumptions (such as connectedness of the range of $h(\theta)$), is equivalent to unbiasedness of t, i.e.

$$E(t \mid \theta) = h(\theta) \qquad \text{for all} \quad \theta . \tag{4}$$

Moreover, from all unbiased estimates t one might wish to find one for which the probability concentration at the true value $h(\theta)$ is maximized for all θ. This would in a similar approximate sense lead to the search for a UMVU estimator. But when, as is usually the case in survey-sampling, the UMVU estimate does not exist, and a prior distribution ξ is available for θ, we may integrate (3) with respect to ξ to obtain

$$P_\xi(|t - h(\theta)| < k) \geqq 1 - \frac{\varepsilon_\xi E((t - h(\theta))^2 \mid \theta)}{k^2} . \tag{5}$$

Here ε_ξ denotes expectation with respect to ξ, and the left-hand side of (5) is the probability of the interval $[t - k, t + k]$ covering the true value $h(\theta)$ in a well-defined random experiment, namely the drawing of θ with distribution ξ followed by the drawing of t (Godambe (1969)). Thus in this situation it would seem natural to find an unbiased estimator t for which the *average* variance $\varepsilon_\xi E((t - h(\theta))^2 \mid \theta)$ is minimum.

Even this recipe of minimising expected variance is recommended only when no more appropriate mode of inference is available. One such mode, for instance, would be the use of the mean of the Bayes posterior distribution of θ obtained on the basis of the data and the prior distribution ξ. This Bayes estimator, however, would usually be very *sensitive* to the variations of the prior distribution ξ; and yet generally the prior knowledge on the part of the statistician is so vague that it can more naturally be characterised by a rather broad *class* C of prior distributions ξ than by any single prior distribution. On the other hand, for the problem of estimation stated in Section 1, there exist classes C of prior distributions large enough to represent adequately the statistician's vague prior knowledge concerning the population yet having the property that there exists a unique unbiased estimate providing the minimum of the expected variance for all $\xi \in C$. This result, with a certain class C of prior distributions, was established by Godambe (1955) and Godambe and Joshi (1965). In the subsequent paragraphs we discuss this in detail, in relation to another class of prior distributions, generalising some related results.

In relation to sampling a finite population, for a *given* sampling design p, an estimator e is said to be an *unbiased* estimator with respect to the population total T in (2) if

$$\textstyle\sum_S e(s, x_i : i \in s)p(s) = T(\mathbf{x}) , \qquad \mathbf{x} \in R_N . \tag{6}$$

For the given p, let B_p denote the class of all unbiased estimators e for the population total T. Further the variance of any unbiased estimator e, $e \in B_p$, is denoted by $V_p(e \mid \mathbf{x})$ where

$$(7) \qquad V_p(e \mid \mathbf{x}) = \sum_S (e(s, x_i : i \in s) - T(\mathbf{x}))^2 p(s)\,, \qquad \mathbf{x} \in R_N\,.$$

Now it is a well-known result (Godambe (1955), Godambe and Joshi (1965)) that except for some very exceptional sampling designs B_p does *not* contain a UMVU estimator e^*. Hence one may proceed to obtain an estimator e^*, $e^* \in B_p$, having a minimum expected variance as described above. Some general results (Godambe (1955), Godambe and Joshi (1965)) in this connection which would be required in the sequel are as follows:

Let α_i, $i = 1, \cdots, N$ be some *given* numbers such that $1 > \alpha_i > 0$ and $\sum_1^N \alpha_i = n$, an integer. Further let C be the class of prior distributions ξ on R_N defined by

$$(8) \qquad C = \left\{\xi : \begin{array}{l} \text{(i) } x_1, \cdots, x_N \text{ when distributed as } \xi \text{ are} \\ \quad\ \text{probabilistically independent and} \\ \text{(ii) } \int x_i\, d\xi = k\alpha_i \text{ where } 0 < k < \infty \end{array}\right\}.$$

Next we define a class D of sampling designs p determined by C as above

$$(9) \qquad D = \{p : \textstyle\sum_{S_i} p(s) = a_i = \alpha_i, i = 1, \cdots, N\}\,,$$

a_i, $i = 1, \cdots, N$ as before denoting the inclusion probabilities (Definition 1.5). Now if we define an estimator e^* as

$$(10) \qquad e^*(s, x_i : i \in s) = \textstyle\sum_{i \in s} x_i/\alpha_i\,,$$

then it is easy to see that $e^* \in B_p$ for all the sampling designs $p \in D$ in (9). In connection with e^* in (10), Godambe and Joshi (1965) proved that the expected variance

$$(11) \qquad \varepsilon_\xi V_p(e^* \mid \mathbf{x}) \leqq \varepsilon_\xi V_p(e \mid \mathbf{x}) \qquad \text{for all} \quad p \in D\,, \quad \xi \in C\,, \quad e \in B_p\,.$$

Further the expected variance $\varepsilon_\xi V_p(e^* \mid \mathbf{x})$ is the *same* for all the sampling designs $p \in D$ in (9) i.e. we may write,

$$(12) \qquad \varepsilon_\xi V_p(e^* \mid \mathbf{x}) = \varepsilon_\xi V_D(e^* \mid \mathbf{x})\,.$$

Now let $\bar{C}$ be the subclass of C in (8) defined by

$$(13) \qquad \bar{C} = \left\{\xi : \begin{array}{l} \text{(i) the same as in (8)} \\ \text{(ii) the same as in (8)} \\ \text{(iii) } \int (x_i/\alpha_i)^2\, d\xi = m^2, i = 1, \cdots, N, 0 < m^2 < \infty \end{array}\right\}.$$

And we enlarge the class of sampling designs D in (9) to, say, $\underline{D}$ by allowing the inclusion probabilities a_i, $i = 1, \cdots, N$ to vary holding the sample size (Definition 1.3), namely $\sum_1^N \alpha_i\ (= n)$, fixed as in D. Thus denoting by $a_i(p)$, $i = 1, \cdots, N$ the inclusion probabilities for the sampling design p and using (1) we write

$$(14) \qquad \underline{D} = \left\{p : \begin{array}{l} \text{(i) } a_i(p) > 0, i = 1, \cdots, N \text{ and} \\ \text{(ii) } \sum_1^N a_i(p) = \sum_1^N \alpha_i \ (= n \text{ fixed}) \end{array}\right\}.$$

With the above notation, (12), (13) and (14), a further result proved by Godambe and Joshi (1965) is that

$$(15) \qquad \varepsilon_\xi V_D(e^* \mid \mathbf{x}) \leqq \varepsilon_\xi V_p(e \mid \mathbf{x})$$

for all $p \in \underline{D}$ in (14), for all estimators $e \in B_p$, and for all $\xi \in \bar{C}$ in (13). In the subsequent sections we shall have occasion to refer to a special case of the inequality (15) obtained by restricting the prior distributions ξ to a subclass $\bar{\bar{C}}$ of $\bar{C}$, where

$$(16) \qquad \bar{\bar{C}} = \left\{ \xi : \begin{array}{l} \text{When } x_i,\ i = 1, \cdots, N \text{ are distributed as } \xi \\ \text{then } x_i/\alpha_i,\ i = 1, \cdots, N \text{ are distributed} \\ \text{independently and identically} \end{array} \right\}.$$

3. Independence and exchangeability. A situation when the prior knowledge is characterizable by the class $\bar{C}$ in (13) of prior distributions may be when the population consists of a number ($= N$ say) of agricultural farms, and the variate value x_i associated with the ith (individual) farm is its produce. Here the statistician may from a previous census possess the knowledge of the acreages (or some proportional numbers α_i ($i = 1, \cdots, N$)) of the different farms. *Without any further prior information*, this prior knowledge may suggest that the prior distribution ξ of the population vector $\mathbf{x} = (x_1, \cdots, x_N)$ (Definition 1.6) is such that for some (unknown) constant of proportionality k, the variates $(x_i - k\alpha_i)/k\alpha_i$, $i = 1, \cdots, N$ have zero means and a common variance. Thus here the prior knowledge is *just enough* to impose some *constraints* on the marginal distributions of the different components x_i, $i = 1, \cdots, N$, for the possible prior distribution ξ. The condition that *there is no more prior knowledge* could be formalised by saying that subject to the above constraints on the marginal distributions the prior distribution ξ should be most *uniform* on R_N. Using the information theory criterion for *uniformity* of distribution, one may note that for *given* marginal distributions, the function $-\int [\xi(\mathbf{x}) \log \xi(\mathbf{x})]\, d\mathbf{x}$ is maximised for the variations of ξ when the joint density ξ is equal to the product of its given marginal densities (Reza (1961)). Hence maximal uniformity is *justification* for the *independence* assumed in the classes C in (8), $\bar{C}$ in (13) and $\bar{\bar{C}}$ in (16) of prior distributions. An analogous result holds for discrete distributions ξ.

In the above justification of independence, we have emphasised the *negative aspect* of our prior knowledge, namely that it does not tell us anything beyond certain constraints to be imposed on the marginals; hence the otherwise *uniformity* of the prior distribution. However, if we emphasise the *positive aspect* of the *same* situation we may require *symmetry* instead of *uniformity* of the prior distribution. For instance in the population of agricultural farms considered in the previous paragraph the prior knowledge may say that the marginal distributions of the variates $(x_i - k\alpha_i)/k\alpha_i$, $i = 1, \cdots, N$ are identical. Here without invoking the argument of uniformity (which as seen before leads to independence) one may emphasise the symmetry, implying that the prior distribution should be *exchangeable* in the variates $(x_i - k\alpha_i)/k\alpha_i$, $i = 1, \cdots, N$. This exchangeability is studied in the subsequent sections.

4. Optimal estimator and design for exchangeable priors. If the statistician's prior knowledge about $\mathbf{x}$ is symmetric in the transformed co-ordinates $y_i = x_i/\alpha_i$, $i = 1, \cdots, N$, he may want to express this prior knowledge in terms of some subclass of the class C^* of priors defined by

$$(17)\qquad C^* = \left\{\xi : \begin{array}{l}\text{When } x_i,\ i = 1, \cdots, N \text{ are distributed as } \xi \\ \text{then } x_i/\alpha_i,\ i = 1, \cdots, N \text{ have an exchangeable} \\ \text{(symmetric) joint distribution}\end{array}\right\}.$$

Clearly the class $\bar{\bar{C}}$ in (16) is a subclass of C^*. Moreover, C^* contains all priors which are obtained by averaging the elements of $\bar{\bar{C}}$ with respect to prior distributions β on $\bar{\bar{C}}$ itself (cf. Ericson (1969)). That is, C^* contains all priors ξ' of the form

$$(18)\qquad \xi'(\mathbf{x}) = \int_{\bar{\bar{C}}} \xi(\mathbf{x})\, d\beta\,.$$

Integrating both sides of (15) with respect to β gives us immediately

$$(19)\qquad \varepsilon_{\xi'} V_D(e^* \mid \mathbf{x}) \leqq \varepsilon_{\xi'} V_p(e \mid \mathbf{x})$$

for all $p \in D$ of (14), whenever $e \in B_p$.

However, not every element of C^* in (17) can be obtained as such a mixture of element of $\bar{\bar{C}}$. (See Hewitt and Savage (1955) for a thorough discussion of exchangeability.) For example, it is easy to see that in general the discrete prior ξ_0 giving equal probability to all permutations of some fixed vector $\mathbf{y} = (y_1, \cdots, y_N)$, $y_i = x_i/\alpha_i$, is not a mixture of independent priors. It is the purpose of the following theorem to extend the result (19) to the rest of the prior distributions in C^*.

THEOREM 4.1. *Let ξ be any element of C^* in* (17), *and let e^* be the estimator defined by* (10). *Then $\varepsilon_\xi V_p(e^* \mid \mathbf{x})$ is the same for all $p \in D$ of* (9); *and if this common value is denoted by $\varepsilon_\xi V_D(e^* \mid \mathbf{x})$ then*

$$(20)\qquad \varepsilon_\xi V_D(e^* \mid \mathbf{x}) \leqq \varepsilon_\xi V_p(e \mid \mathbf{x})$$

for all $p \in \underline{D}$ of (14) *whenever $e \in B_p$.*

PROOF. It is sufficient to prove (20) for any prior ξ_0 giving equal probability to all permutations of the co-ordinates of some fixed vector $\mathbf{y} = (y_1, \cdots, y_N)$ where $y_i = x_i/\alpha_i$; then (20) will follow for general $\xi \in C^*$ by integration.

Now let e be an unbiased estimator for the population total, given a sampling design $p \in \underline{D}$ in (14). Then

$$\textstyle\sum_s p(s)e(s, x_i : i \in s) = \sum_1^N x_i \qquad \text{for all} \quad \mathbf{x} \in R_N\,.$$

Putting $x_i = \alpha_i y_i$, $i = 1, \cdots, N$ and $e(s, x_i : i \in s) = f(s, y_i : i \in s)$ gives

$$(21)\qquad \textstyle\sum_s p(s)f(s, y_i : i \in s) = \sum_1^N \alpha_i y_i \qquad \text{for all} \quad \mathbf{y} \in R_N\,.$$

The right-hand side of (20) is obtained by taking the average of the variance of $f(s, y_i : i \in s)$ over all $N!$ vectors $\mathbf{y}_\pi = (y_{\pi(1)}, \cdots, y_{\pi(N)})$ (not necessarily distinct) obtained from our fixed vector $\mathbf{y}$ by permutations π of the co-ordinate indices. That is,

$$(22)\qquad \varepsilon_{\xi_0} V_p(e\,|\,\mathbf{x}) = \frac{1}{N!} \sum_\pi [\sum_s p(s) f^2(s, y_{\pi(i)} : i \in s) - (\sum_1^N \alpha_i y_i)^2]$$

$$= \frac{1}{N!} \sum_\pi [\sum_s p(s) f^2(s, y_{\pi(i)} : i \in s)] - \frac{1}{N!} \sum_\pi (\sum_1^N \alpha_i y_i)^2 .$$

Taking the same kind of average of both sides of (21), we have

$$(23)\qquad \frac{1}{N!} \sum_\pi \sum_s p(s) f(s, y_{\pi(i)} : i \in s) = \frac{1}{N!} \sum_\pi \sum_1^N \alpha_i y_{\pi(i)}$$

$$= \frac{(N-1)!}{N!} \sum_1^N \alpha_i \sum_1^N y_i .$$

Noting that $\sum_1^N \alpha_i = n$, and putting $\bar{y} = \sum_1^N y_i/N$ we may write (23) more simply as

$$(24)\qquad \sum_{\pi,s} \frac{p(s)}{N!} f(s, y_{\pi(i)} : i \in s) = n\bar{y} .$$

Now from (24) we see that $f(s, y_{\pi(i)} : i \in s)$ is an unbiased 'estimator' of $n\bar{y}$ if the sampling design consists of drawing the sample s from $\{1, \cdots, N\}$ by the sampling design p and observing $y_{\pi(i)} : i \in s$ for a randomly chosen permutation π of the vector $\mathbf{y}$. We may denote this sampling design by (p, π).

(REMARK. This *usage* of the word 'estimator' is extended somewhat outside the restrictions of Definition 1.7, in the sense that the value of f depends in addition to s on the particular permutation π chosen at random. While this *extended usage* helps in accomplishing the proof of Theorem 4.1, the discussion in Section 5 of Godambe and Joshi (1965), shows that otherwise restriction to Definition 1.7 does not in any way affect the generality of results in the paper.)

The variance of the estimator is evidently

$$(25)\qquad \sum_{\pi,s} \frac{p(s)}{N!} f^2(s, y_{\pi(i)} : i \in s) - (n\bar{y})^2 .$$

Suppose for the moment that the co-ordinates y_i in the fixed vector $\mathbf{y}$ are all distinct. (A similar argument can be carried through if they are not.) The sum $\sum_{\pi,s}$ in (25) contains $N!\,{}^NC_n$ terms, since for each of the $N!$ permutations π we sum over ${}^NC_n$ samples s. If we consider corresponding to each term the set $\mathbf{y}^j$ of the unlabelled (i.e. without the individual index $\pi(i)$) values $[y_{\pi(i)} : i \in s]$, the number of distinct sets $\mathbf{y}^j$ is ${}^NC_n$, so that we may regard the index j as running from 1 to ${}^NC_n$, and each $\mathbf{y}^j$ is the set of values used in $N!$ terms of $\sum_{\pi,s}$. Now consider the estimator $g(s, y_{\pi(i)} : i \in s)$ obtained as the conditional expectation of $f(s, y_{\pi(i)} : i \in s)$ with respect to the sampling design (p, π), given that the unlabelled values $[y_{\pi(i)} : i \in s]$ are $\mathbf{y}^j$. That is, if

$$\sum_s^{j \text{ fixed}} \sum_\pi^{s,j \text{ fixed}} \quad \text{means} \quad \sum_s \sum_{\{\pi : [y_{\pi(i)} : i \in s] = \mathbf{y}^j\}}$$

then

$$g(\mathbf{y}^j) \textstyle\sum_s^{j \text{ fixed}} \sum_\pi^{s,j \text{ fixed}} \left[\frac{p(s)}{N!}\right] = \sum_s^{j \text{ fixed}} \sum_\pi^{s,j \text{ fixed}} \left[\frac{p(s)}{N!}\right] f(s, y_{\pi(i)} : i \in s) . \tag{26}$$

Certainly, $g(s, y_{\pi(i)} : i \in s)$, which may be written as $g(\mathbf{y}^j)$, is an unbiased estimator of $n\bar{y}$ with respect to (p, π): and its variance must be no greater than that of f, by an argument similar to the one used in proving the Rao–Blackwell theorem. Hence

$$\textstyle\sum_{\pi,s} \frac{p(s)}{N!} f^2(s, y_{\pi(i)} : i \in s) \geqq \sum_{\pi,s} \frac{p(s)}{N!} g^2(\mathbf{y}^j) . \tag{27}$$

Moreover, it follows from the fact that the "order statistic" $\mathbf{y}^j$ is complete (Royall (1968)) that if (21) is to hold for all $\mathbf{y} \in R_N$ then $g(\mathbf{y}^j)$ must in fact be the sample total $\sum_{i \in s} y_{\pi(i)}$, the sum of all elements of $\mathbf{y}^j$. Hence

$$\textstyle\sum_{\pi,s} \frac{p(s)}{N!} [\sum_{i \in s} y_{\pi(i)}]^2 \leqq \sum_{\pi,s} \frac{p(s)}{N!} f^2(s, y_{\pi(i)} : i \in s) . \tag{28}$$

But the left-hand side of (28) may be replaced by

$$\textstyle\sum_{\pi,s} \frac{p'(s)}{N!} [\sum_{i \in s} y_{\pi(i)}]^2 \tag{29}$$

for *any* sampling design p', in particular a design with size n and inclusion probabilities $a_i = \alpha_i$ (Definition 2.5). Subtracting $(1/N!) \sum_\pi [\sum_1^N \alpha_i y_i]^2$ from each side of (28) and using (22) gives

$$\textstyle\sum_{\pi,s} \frac{p'(s)}{N!} [\sum_{i \in s} y_{\pi(i)}]^2 - \frac{1}{N!} \sum_\pi [\sum_1^N \alpha_i y_i]^2 \leqq \varepsilon_{\xi_0} V_p(e \mid \mathbf{x}) . \tag{30}$$

Finally, the left-hand side of (30) is just $\varepsilon_{\xi_0} V_{p'}(e^* \mid \mathbf{x})$ for e^* defined by (10) and any design $p' \in D$ of (9); explicitly, it is

$$\frac{n}{N} \textstyle\sum_1^N (x_i/\alpha_i)^2 + \frac{n(n-1)}{N(N-1)} \sum_{i \neq i'} x_i x_{i'}/\alpha_i \alpha_{i'} - \frac{1}{N!} \sum_\pi [\sum_1^N x_i]^2 . \tag{31}$$

Thus the theorem is proved.

In practical terms, what we have shown is the following: If the statistician's prior knowledge can be approximated by C^* and his resources limit him to taking a sample of size n, then for estimation of the population total it is advisable for him to select a design from the class D in (9) and the estimator e^* of (10).

Some very special cases of this result have previously been established. The proof that

$$\varepsilon_\xi V_p(e^* \mid \mathbf{x}) \leqq \varepsilon_\xi V_p(e \mid \mathbf{x}) \tag{32}$$

was given by Kempthorne (1968) under the conditions that p (on both sides of (32)) is simple random sampling without replacement with n draws; $e^*(s, x_i : i \in s)$ is $(N/n) \sum_{i \in s} x_i$; the prior ξ is exchangeable with respect to the x_i; and e is any estimator which is homogeneous linear in the sampled x_i, 'origin invariant' and unbiased with respect to p. Rao (1971) proved (32) under the more general conditions that p is a sampling design with inclusion probabilities α_i; $e^*(s, x_i : i \in s)$ is $\sum_{i \in s} x_i/\alpha_i$; the distribution ξ is exchangeable in the variates x_i/α_i; and e is homogeneous linear and unbiased with respect to p. Thompson (1971) showed that the assumption of linearity of e could be dropped. The Theorem 6.1 contains all these results, and establishes the optimality of a *sampling design* with inclusion probabilities α_i as well as optimality of the estimator $\sum_{i \in s} x_i/\alpha_i$.

5. Necessity of fixed sample size design. In Section 4 we have proved joint optimality of an estimator and a sampling design under the assumption that all competing designs have (the same) *fixed* sample size. This assumption is used in an essential way at the point where completeness of the order statistic $\mathbf{y}^j$ is invoked. The following example illustrates the fact that if variable sample sizes are allowed in the design with inclusion probabilities α_i, the estimator e^* of (10) need not have minimal expected variance with respect to a prior which is exchangeable in the x_i/α_i.

Let the population size N be 3, and suppose that a priori the variates $y_1 = x_1$, $y_2 = \frac{3}{2}x_2$ and $y_3 = 3x_3$ are exchangeable. Then the variable sample size design defined by

$$p(\{1\}) = p(\{1, 2\}) = p(\{1, 2, 3\}) = \tfrac{1}{3}$$

has inclusion probabilities $\alpha_1 = 1$, $\alpha_2 = \frac{2}{3}$, $\alpha_3 = \frac{1}{3}$. The Horvitz–Thompson estimator, which corresponds to e^* in this situation, is given by

$$f^*(\{1\}; y_1) = y_1,\ f^*(\{1, 2\}; y_1, y_2) = y_1 + y_2\,,$$
$$f^*(\{1, 2, 3\}; y_1, y_2, y_3) = y_1 + y_2 + y_3\,.$$

The average of its variance over all permutations of $\{y_1, y_2, y_3\}$ is

$$V_{HT} = \tfrac{4}{27}(y_1^2 + y_2^2 + y_3^2) + \tfrac{2}{27}(y_1y_2 + y_1y_3 + y_2y_3)\,.$$

Now consider the alternative estimator

$$f(\{1\}; y_1) = y_1,\ f(\{1, 2\}; y_1, y_2) = y_1 + \tfrac{2}{3}y_2\,,$$
$$f(\{1, 2, 3\}; y_1, y_2, y_3) = y_1 + \tfrac{4}{3}y_2 + y_3\,.$$

This has *expected* variance

$$\tfrac{14}{81}(y_1^2 + y_2^2 + y_3^2) + \tfrac{4}{27}(y_1y_2 + y_1y_3 + y_2y_3)\,,$$

which is sometimes less than V_{HT}, for example when $\mathbf{y} = (1, -1, 0)$. Thus for this sampling design, which has the proper inclusion probabilities as determined by the prior knowledge, the Horvitz–Thompson estimator does not always have least expected variance.

REFERENCES

ERICSON, W. A. (1969). Subjective Bayesian models in sampling finite populations (with discussion). *J. Roy. Statist. Soc. Ser. B.* **31** 195–233.

GODAMBE, V. P. (1955). A unified theory of sampling from finite populations. *J. Roy. Statist. Soc. Ser. B.* **17** 267–278.

GODAMBE, V. P. (1965). A review of the contributions towards a unified theory of sampling from finite populations. *Internat. Statist. Inst. Rev.* **33** 242–258.

GODAMBE, V. P. (1969). Some aspects of theoretical developments in survey-sampling. *New Developments in Survey-Sampling*. Wiley, New York. 27–58.

GODAMBE, V. P. (1969). A law of large numbers for sampling finite populations with different inclusion probabilities for different individuals (abstract). *Ann. Math. Statist.* **40** 2218–2219.

GODAMBE, V. P. and JOSHI, V. M. (1965). Admissibility and Bayes estimation in sampling from finite populations—I. *Ann. Math. Statist.* **36** 1707–1722.

HEWITT, E. and SAVAGE, L. J. (1955). Symmetric measures on Cartesian products. *Trans. Amer. Math. Soc.* **80** 470–501.

KEMPTHORNE, O. (1969). Some remarks on statistical inference in finite sampling. *New Developments in Survey-Sampling*, (N. L. Johnson and H. Smith, eds.). Wiley, New York. 671–695.

RAO, C. R. (1971). Some aspects of statistical inference in problems of sampling from finite populations. *Foundations of Statistical Inference*, (V. P. Godambe and D. A. Sprott, eds.). Holt, Rinehart & Winston, Toronto. 177–190.

REZA, J. M. (1961). *An Introduction to Information Theory*. McGraw-Hill, New York. 276.

ROYALL, R. (1968). An old approach to finite population sampling theory. *J. Amer. Statist. Assoc.* **63** 1269–1279.

THOMPSON, M. E. (1971). *Foundations of Statistical Inference*. (V. P. Godambe and D. A. Sprott, eds.). Holt, Rinehart and Winston Toronto. 196–198.

DEPARTMENT OF STATISTICS
UNIVERSITY OF WATERLOO
WATERLOO, ONTARIO, CANADA