The Annals of Mathematical Statistics
1970, Vol. 41, No. 5, 1774–1779

FINITE POPULATION SAMPLING—ON LABELS IN ESTIMATION

By Richard M. Royall

The Johns Hopkins University

1. Introduction. This paper is concerned with finite, labelled populations—i.e., with each unit in the population is associated a number, which is unknown and is of interest, and a unique label, which has been assigned by some, possibly unknown, procedure. Some results of an investigation of the role of such labels in sampling and inference are presented. For an arbitrary fixed sampling plan, certain aspects of the use of unit labels in estimation are examined. The estimators of Horvitz and Thompson [7] are familiar practical examples of estimators which depend, not only on the numbers observed, but also on their associated labels; artificial examples of estimators which depend on the labels are occasionally produced as counterexamples to claims of optimality properties for certain popular estimators (See, e.g., Roy and Chakravarti [10]).

Here it is shown (Theorem, Section 4) that for general sampling plans and for many parameters of interest, the class of estimators which do not depend on the labels identifying the units in the sample has a certain property which seems desirable when little is known about the relation between the number and the label associated with each unit. In particular, a theoretical (minimax) justification is given (Corollary 1) for the common practice of ignoring the labels when estimating from simple random samples. These results are then applied (Section 5) to general linear unbiased estimators of the population mean, and it is shown that for the case of simple random sampling, with the parameter space subject to certain natural restrictions, the sample mean is minimax (convex loss function) among linear unbiased estimators.

2. Description of the problem. The population of interest can be described as

(i) a set of N distinct units, together with
(ii) a set of N real numbers, one associated with each unit, and
(iii) a set of labels, say the integers $1, 2, \cdots, N$, which identify the units.

The problem to be considered is that of estimating the value of a real-valued symmetric function $\theta(\mathbf{x})$ of the components of the parameter vector $\mathbf{x} = (x_1, x_2, \cdots, x_N)$ whose ith element is the number associated with the unit labelled "i." Let $\pi(1), \pi(2), \cdots, \pi(N)$ be a permutation of the integers $1, 2, \cdots, N$. If the units are relabelled so that "i" now identifies the unit originally labelled "$\pi(i)$," then the parameter vector becomes $x_\pi = (x_{\pi(1)}, x_{\pi(2)}, \cdots, x_{\pi(N)})$, and the quantity to be estimated is $\theta(\mathbf{x}_\pi)$. By symmetry $\theta(\mathbf{x}) = \theta(\mathbf{x}_\pi)$ for all permutations π; i.e. θ is invariant under relabelling.

Let S denote the collection of all subsets, s, (distinct elements) of the set of labels $\{1, 2, \cdots, N\}$. If $p(\cdot)$ is a probability function on S, then the sampling rule

Received September 14, 1967; revised April 7, 1970.

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

is "Observe $\{(i, x_i), i \in s\}$ with probability $p(s)$." See Godambe [6] for sufficiency of $\{(i, x_i), i \in s\}$ under various schemes for accomplishing the actual selection of the sample in accordance with the probability function p. Denote by $n(s)$ the number of integers in s and by $P(n)$ the probability, $\sum_{\{s';\, n(s')=n\}} p(s')$, that the sample contains exactly n(different) units.

Any function $t(s, \mathbf{x})$ which depends on $\mathbf{x}$ only through those coordinates x_i for which i is in s (i.e., does not depend on the coordinates not observed) can be used as an estimator for θ.

The pair $p:t$ where p is a sampling plan and t is an estimator will be called a strategy. A decision-theoretic approach to the evaluation of strategies will be taken. Thus it is assumed that when θ is estimated to be a number b, a loss of $l(b, \theta)$ units is incurred. For a strategy $p:t$ and a parameter vector $\mathbf{x}$, the expected loss, or risk, denoted by $R(t, p, \mathbf{x})$ is

$$R(t, p, \mathbf{x}) = \sum_{s \in S} p(s)\, l(t(s, \mathbf{x}), \theta(\mathbf{x})). \tag{1}$$

Throughout this work the N units and their associated numbers are treated as being fixed. There are $N!$ ways in which the units might be labelled. If $\mathbf{x}^*$ is a vector whose components are the N numbers arranged in non-decreasing order, then the parameter vector $\mathbf{x}$ is one of the (at most) $N!$ vectors obtained by permuting the components of $\mathbf{x}^*$, and the quantity to be estimated is $\theta(\mathbf{x}^*)$. The average risk over all permutations of the labels is

$$\bar{R}(t, p, \mathbf{x}^*) = \sum_{\pi} R(t, p, \mathbf{x}_\pi)/N!. \tag{2}$$

It is assumed that the labels serve only to identify the units, and that there is no available knowledge of any systematic relationship between the labels and the x's. This assumption might be expressed probabilistically as: for fixed $\mathbf{x}^*$ and for every permutation π

$$\Pr(\mathbf{x} = \mathbf{x}_\pi^*) = 1/N!. \tag{3}$$

Equation (3) would be satisfied if the numbers $x_1, \cdots, x_N$ were realized values of exchangeable random variables $X_1, \cdots, X_N$ or if the units were labelled at random. More frequently, however, the assumption is met only in terms of personal, subjective probabilities—in populations where "the labels are uninformative." When (3) holds the average risk (2) can be interpreted as a "Bayes risk" with respect to the a priori probability distribution under which each permutation of $\mathbf{x}^*$ is equally probable as the value of $\mathbf{x}$. Note that the present model is not a complete Bayesian model—the vector $\mathbf{x}^*$ is fixed, but unknown, and it is assigned no a priori probability distribution. Ericson [3] has examined finite population sampling theory under a complete Bayesian model. The present paper is an attempt to gain further understanding of some situations in which "labels are uninformative" or "labels are irrelevant" under weak "a priori" assumptions. A population which is not of the type considered here is a group of N college freshmen classified (stratified) according to sex with x_i the number of cigarettes smoked per day by the ith student. If the boys are identified by labels $1, 2, \cdots, N_1$, and the girls by

$N_1+1, N_1+2, \cdots, N$, then the labels are "informative"—e.g., if one observation is (2, 10) then the label 2 indicates that this student who smokes 10 cigarettes per day is a male, which, assuming a sex difference in smoking habits, is a relevant bit of information. This population also furnishes a simple example of an important parametric function which does not satisfy the symmetry condition: the difference between the average cigarette consumption of freshman boys and girls is

$$\theta_d(\mathbf{x}) = \sum_{i=1}^{N_1} x_i/N_1 - \sum_{i=N_1+1}^{N} x_i/N - N_1,$$

which is not, in general, equal to $\theta_d(\mathbf{x}_\pi)$ for arbitrary π.

3. Symmetric estimators. If data $\{(i, x_i); i \in s\}$ are observed, let y_i, $i = 1, 2, \cdots, n(s)$, denote the ith smallest among the observed numbers $\{x_i; i \in s\}$. An estimator $t(s, \mathbf{x})$ will be called symmetric if it depends on the data only through $n(s)$ and the "order statistic" $y_1, y_2, \cdots, y_{n(s)}$. A symmetric estimator is one determined by the numbers associated with the units selected in the sample and not by the labels which identify these units. The sample mean

$$t(s, \mathbf{x}) = \sum_{i \in s} x_i/n(s) = \sum_{i=1}^{n(s)} y_i/n(s)$$

is symmetric.

Blackwell and Girshick [2], pages 229–233, showed that, if t is any symmetric estimator and p is any fixed sample size sampling plan say $P(n) = 1$, then

$$\max_\pi R(t, p, \mathbf{x}_\pi) \geqq \max_\pi R(t, p^*, \mathbf{x}_\pi), \tag{4}$$

where p^* is a simple random sampling plan—

$$p^*(s) = 1/\binom{N}{n} \quad \text{if} \quad n(s) = n;$$
$$= 0 \quad \text{otherwise.}$$

This result says that, when a symmetric estimator is to be used, p^* is always a minimax sampling plan.

Frequently, for administrative or other reasons, a sampling plan other than p^* is used. An example of this is a survey in which the primary objective is to investigate a characteristic y, and for meeting this objective a stratified sampling plan is used. However, the characteristic x, which is not believed to be related to y or to the stratification criterion, is also of interest. Rao [9] has investigated a problem of this general sort. In such problems, for estimating the total, $T = \sum_{i=1}^{N} x_i$, the (generally non-symmetric) unbiased estimator of Horvitz and Thompson [7] is frequently used. This estimator is $t_{HT} = \sum_{i \in s} x_i/q(i)$, where $q(i)$ is the probability $\sum_{\{s;\, i \in s\}} p(s)$ that unit i is in the sample. The estimator is used despite the intuitive feeling that the simple fact "unit i had smaller probability of being selected in the sample than unit j had" is, *of itself*, quite a weak excuse for attaching a larger weight to x_i than to x_j in estimating T. The next section, in which certain properties of symmetric estimators are studied for arbitrary fixed sampling plan, p, as well as for p^*, supports this intuition.

4. Average risk of symmetric estimators. For any sampling plan p and estimator t, a corresponding symmetric estimator can be produced by a procedure of averaging: if $p(s) > 0$ let

$$\bar{t}(s, \mathbf{x}) = \sum_\pi t(\pi^{-1}(s), \mathbf{x}_\pi)p(\pi^{-1}(s))/(n(s))!(N-n(s))!P(n(s)) \tag{5}$$

where $\pi^{-1}(s) = \{j; \pi(j) \in s\}$. Define $\bar{t}(s', x) = 0$ if $p(s') = 0$.

Here $t(\pi^{-1}(s), \mathbf{x}_\pi)$ is the value the estimator t would assume if the population were relabelled, the unit originally labelled "$\pi(i)$" now bearing the label "i," and the same set of *units* were selected in the sample. (Of course $t(\pi^{-1}(s), \mathbf{x}_\pi) \neq t(s, \mathbf{x})$ in general.) The probability of selection of this set of units, which now is identified by the labels in $\pi^{-1}(s)$, is $p(\pi^{-1}(s))$. *Note that $t(s, \mathbf{x})$ is symmetric—it depends on the data $\{(i, x_i); i \in s\}$ only through $n(s)$ and the "order statistic" $y_1, y_2, \cdots, y_{n(s)}$.* Thus $\bar{t}(\pi^{-1}(s), \mathbf{x}_\pi) = t(s, \mathbf{x})$ for all permutations π.

The following theorem shows that, in terms of average (over labelling systems) risk (2), $\bar{t}$ is a better estimator than t.

THEOREM. *If the loss function, $l(a, \theta)$, is convex in the first argument, then*

$$\bar{R}(\bar{t}, p, \mathbf{x}^*) \leqq \bar{R}(t, p, \mathbf{x}^*). \tag{6}$$

PROOF. Recall that $\theta(\mathbf{x}_\pi) = \theta(\mathbf{x}^*)$ for all π.

$$\begin{aligned}
N!\bar{R}(\bar{t}, p, \mathbf{x}^*) &= \sum_{\pi'} \sum_s l\Big(\sum_\pi t(\pi^{-1}(s), (\mathbf{x}_{\pi'})_\pi)p(\pi^{-1}(s))/n(s)!(N-n(s))! \\
&\qquad \cdot P(n(s)), \theta(\mathbf{x}^*)\Big)p(s) \\
&\leqq \sum_{\pi'} \sum_s \sum_\pi l(t(\pi^{-1}(s), (\mathbf{x}_{\pi'})_\pi), \theta(\mathbf{x}^*))p(\pi^{-1}(s))p(s)/(n(s))! \\
&\qquad \cdot (N-n(s))!P(n(s)) \\
&= \sum_{\pi'} \sum_s \sum_\pi l(t(\pi^{-1}(s), \mathbf{x}_{\pi'}), \theta(\mathbf{x}^*))p(\pi^{-1}(s))p(s)/(n(s))! \\
&\qquad \cdot (N-n(s))!P(n(s)) \\
&= \sum_{\pi'} \sum_s \sum_{\{s';n(s')=n(s)\}} l(t(s', \mathbf{x}_{\pi'}), \theta(\mathbf{x}^*))p(s')p(s)/P(n(s)) \\
&= \sum_{\pi'} \sum_{n=1}^N \sum_{\{s';n(s')=n\}} l(t(s', \mathbf{x}_{\pi'}), \theta(\mathbf{x}^*)p(s') \\
&= \sum_{\pi'} \sum_{s'} l(t(s', \mathbf{x}_{\pi'}), \theta(\mathbf{x}^*))p(s') \\
&= \sum_{\pi'} R(t, p, \mathbf{x}_{\pi'}). \qquad \square
\end{aligned}$$

For a strictly convex loss function, the inequality in (6) is strict unless $\bar{t}(s, \mathbf{x}_\pi) = t(s, \mathbf{x}_\pi)$ for all s in S for which $p(s) > 0$ and all permutations π.

In case $p(s) = p(s')$ for all s, s' having $n(s) = n(s')$, expression (5) reduces to

$$\bar{t}(s, \mathbf{x}) = \sum_\pi t(\pi^{-1}(s), \mathbf{x}_\pi)/N! \tag{7}$$

Since in this case $R(\bar{t}, p, \mathbf{x}_\pi)$ does not depend on the permutation π, the following is immediate from the theorem:

COROLLARY 1. *If $p(s) = p(s')$ for all s, s' having $n(s) = n(s')$ then*

$$\max_\pi R(\bar{t}, p, \mathbf{x}_\pi) \leqq \max_\pi R(t, p, \mathbf{x}_\pi) \tag{8}$$

with strict inequality unless $\bar{t}(s, \mathbf{x}_\pi) = t(s, \mathbf{x}_\pi)$ for all s with $p(s) > 0$ and all π.

Thus for a simple random sample of n units without replacement, unless t depends on the data $\{(i, x_i), i \in s\}$ only through the "order-statistic" $y_1, y_2, \cdots, y_n$, there is another estimator such that (i) it is a function only of this "order-statistic," (ii) its risk function depends only on $\mathbf{x}^*$ and not on the way in which the units are labelled, and (iii) unless $\bar{t}(s, \mathbf{x}_\pi) = t(s, \mathbf{x}_\pi)$ for all π and all s containing n elements, $\bar{t}$ has smaller average (and smaller maximum) risk over the $N!$ parameter points obtained by rearranging the labels on the N numbers $x_1{}^*, x_2{}^*, \cdots, x_N{}^*$.

5. Symmetric estimators for the population total. Let $\theta(\mathbf{x})$ be the population total, $T = \sum_{i=1}^N x_i$, and let t_L be any linear unbiased estimator for T, i.e. (Godambe [3]) $t_L(s, \mathbf{x}) = \sum_{i=1}^N b_i(s) x_i$ where $b_i(s) = 0$ if $i \notin s$, and $\sum_{\{s; i \in s\}} b_i(s) p(s) = 1$. In this case

$$\bar{t}_L(s, \mathbf{x}) = \frac{1}{n(s)} \left(\sum_{i \in s} x_i\right) \sum_{k=1}^N {}^* b_k(s') p(s') / P(n(s)) \tag{9}$$

where $\sum_k^*$ denotes summation over the set $\{s'; k \in s' \text{ and } n(s') = n(s)\}$. If the sample size is fixed, say at n, and if $t_0(s, \mathbf{x})$ denotes $N \sum_{i \in s} x_i / n(s)$, then (9) reduces to

$$t_L(s, \mathbf{x}) = t_0(s, \mathbf{x}). \tag{10}$$

In particular (10) applies when stratified random sampling is performed and t_L is the conventional unbiased estimator.

Thus the theorem yields

COROLLARY 2. *If $P(n) = 1$ for some n (fixed sample size) and if $t_L(s, \mathbf{x})$ is any linear unbiased estimator for T, then*

$$\bar{R}(t_0, p, \mathbf{x}^*) \leqq \bar{R}(t_L, p, \mathbf{x}^*). \tag{11}$$

COROLLARY 3. *For the simple random sampling plan p^*, if t_L is any linear unbiased estimator for T, then*

$$\max_\pi R(t_0, p^*, \mathbf{x}_\pi) \leqq \max_\pi R(t_L, p^*, \mathbf{x}_\pi). \tag{12}$$

Corollary 2 and Corollary 3 imply that, for simple random sampling of n units without replacement, and for any convex loss function, the sample mean $\sum_{i \in s} x_i / n$ is a best linear unbiased estimator of T/N, in the sense of minimizing both the average and the maximum risk over the $N!$ parameter points obtained by relabelling the units in the population.

For the case of simple random sampling Aggarwal [1] proved that the sample mean is a minimax estimator for T/N (not just minimax among linear unbiased estimators) if the parameter space is given by $\{\mathbf{x}; \sum_{i=1}^N (x_i - (T/N))^2 \leqq \text{const.}\}$ and $l(b, \theta) = (b, \theta)^2$. At least one attempt has been made (Joshi [8]) to extend Aggarwal's result to arbitrary "symmetrical" parameter spaces. But it is well known (Ferguson

[4] page 92) that the sample mean is not minimax (unless $n = N$) over the parameter space composed of the 2^N vectors in which each component is either $+1$ or -1. This parameter space is "symmetrical" according to most definitions which the author can imagine, but if $N > 1$ and $n = 1$, then the estimator which takes the value $y/2$ when y is observed has smaller maximum risk than that of the sample mean, y. The results of the present section, however, apply to any parameter space which is closed under relabelling (permutation) of coordinates.

REFERENCES

[1] AGGARWAL, O. P. (1959). Bayes and minimax procedures in sampling from finite and infinite populations I. *Ann. Math. Statist.* **30** 206–218.

[2] BLACKWELL, D. and GIRSHICK, M. A. (1954). *Theory of Games and Statistical Decisions*, Wiley, New York.

[3] ERICSON, W. A. (1969). Subjective Bayesian models in sampling finite populations. *J. Roy. Statist. Soc. Ser. B* **31** 195–224.

[4] FERGUSON, T. S. (1967). *Mathematical Statistics: A Decision Theoretic Approach.* Academic Press, New York.

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

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

[7] HORVITZ, D. G. and THOMPSON, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* **47** 663–685.

[8] JOSHI, V. M. (1966). Admissibility and Bayes estimation in sampling finite populations IV. *Ann. Math. Statist.* **37** 1658–1670.

[9] RAO, J. N. K. (1966). Alternative estimators in pps sampling for multiple characteristics. *Sankhyā Ser. A* **28** 47–60.

[10] ROY, J. and CHAKRAVARTI, I. M. (1960). Estimating the mean of a finite population. *Ann. Math. Statist.* **31** 392–398.