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.