BUREAU OF THE CENSUS
STATISTICAL RESEARCH DIVISION
Statistical Research Report Series
No. RR90/03
*Previously TN90/01

Formulas for Approximating Overlap Probability When Maximizing Overlap

by

Lawrence R. Ernst

U. S. Bureau of the Census
Statistical Research Division
Washington D.C. 20233

Report Issued 7/23/90

This series contains research reports, written by or in cooperation with, staff members of the Statistical Research Division, whose content may be of interest to the general statistical research community. The views reflected in these reports are not necessarily those of the Census Bureau nor do they necessarily represent Census Bureau statistical policy or practice. Inquiries may be addressed to the author(s) or the SRD Report Series Coordinator, Statistical Research Division, Bureau of the Census, Washington D.C. 20233.

1. INTRODUCTION

This report is not intended as a stand-alone document. It is assumed that the reader has a basic understanding of Ernst (1986). Potential readers who do not, proceed at their own risk.

It has been proposed that the expected overlap of sample PSUs be used as a criterion for selection of a stratification in the redesign for those household surveys that will maximize overlap of PSUs using the procedure of Ernst (1986). In general, computing the exact expected overlap for several stratifications that are being evaluated would require the impractical operation of running the entire overlap procedure for each stratum, for each stratification.

A simple formula that provides a reasonable approximation to the overlap probability for each stratum in the new design would be a practical alternative. Such a formula, (2.4), is presented in this report. This formula requires that the initial and new stratifications be known, as well as the probability of selection for each PSU in the new design. (2.4) is an exact expression for the overlap probability when the new selection probability for each PSU in a new design stratum is equal to its initial selection probability. This formula will generally provide a good approximation to the overlap probability if the selection probabilities for the two designs are reasonably close.

2. THE APPROXIMATION FORMULA

Although the procedure of Ernst (1986) is presented in much more generality, the work here will be limited to the case when both the initial and new designs are one PSU per stratum. All surveys for which the use of this overlap procedure is under consideration satisfy this case.

Some notation will first be introduced. The notation of Ernst (1986) will be used in a modified form arising from the simplifications possible in the one PSU per stratum case.

Let S denote a stratum in the new design and let r denote the number of strata in the initial design containing PSUs in S. For $i=1,\ldots,r$, let u_i denote the number of PSUs in S that are in the i-th initial stratum, and then for $j=1,\ldots,u_i$, let p_{ij}, π_{ij} denote the probability of selection in the initial and new designs, respectively, of the j-th PSU from the i-th initial stratum that is in S. Finally, for $i=1,\ldots,r$, let y_i denote the probability that the i-th initial stratum is selected as the initial stratum on whose sample PSUs the selection probabilities are conditioned.

In general, an upper bound on the probability of overlap when using the procedure of Ernst (1986) is

$$\sum_{i=1}^{r}\sum_{j=1}^{u_i}([\min\{\pi_{ij},y_ip_{ij}\}] + [p_{ij}(\pi_{ij}-\min\{\pi_{ij},y_ip_{ij}\})]) \tag{2.1}$$

where the y_i's are set which maximize (2.1). Furthermore,

> (2.1) is an exact expression for the overlap probability if and only if for all i,j the ij-th PSU is never in the new sample when the i-th initial stratum is selected and the ij-th PSU was not in the initial sample. (2.2)

In order to obtain a set of y_i's which maximize (2.1), and thus be able to express (2.1) in terms of PSU selection probabilities only, it is generally necessary to run the linear programming procedure in Ernst (1986). However, as will be shown, in the special case when $p_{ij}=\pi_{ij}$ for all i,j, an optimal set of y_i's is easily obtained, leading to (2.4), and furthermore, (2.2) holds, so that (2.4) is an exact expression for the overlap probability.

First, however, it will be explained how (2.1) and (2.2) were derived. For each i,j, there are three cases in which the ij-th PSU can be selected in the new sample, namely:

(a) The i-th initial stratum is the selected stratum and the ij-th PSU was in the initial sample.

(b) The i-th initial stratum is not the selected stratum.

(c) The i-th initial stratum is the selected stratum and the ij-th PSU was not in the initial sample.

The probabilities that the ij-th PSU was in the initial sample for cases (a), (b) and (c) are 1, p_{ij} and 0, respectively. (For (b), this probability is not necessarily p_{ij} if the sampling was not independent from stratum to stratum in the initial design, but p_{ij} is used anyway for lack of a better value. This point is discussed in more detail in Ernst (1986).) Consequently, the optimization procedure seeks to maximize the selection of the ij-th PSU when (a) occurs, and to minimize its selection when (c) occurs.

Consider now the terms in (2.1). The term within the first set of brackets is the joint probability that (a) occurs and that the ij-th PSU is selected in the new sample. This is because $y_i p_{ij}$ is the probability of (a) occurring and the ij-th PSU will always be selected in this case unless $y_i p_{ij}$ exceeds π_{ij}.

The term within the second set of brackets is an upper bound on the joint probability that either (b) or (c) occurs and the ij-th PSU is in both samples, and is an exact value if and only if this PSU is never selected in the new sample when (c) occurs. This is because $\pi_{ij} - \min\{\pi_{ij}, y_i p_{ij}\}$ is the probability that (b) or (c) occurs and the ij-th PSU is in the new sample. In this event, the conditional probability that this PSU was also in the initial sample is p_{ij} if (c) doesn't occur when this PSU is in the new sample, but is strictly less than p_{ij} otherwise. Thus (2.1) and (2.2) are established.

Returning to the question of finding a set of y_i's which maximize (2.1) when $p_{ij} = \pi_{ij}$ for all i,j, first note that in this case (2.1) reduces to

$$\sum_{i=1}^{r} \sum_{j=1}^{u_i} [(\pi_{ij} - \pi_{ij}^2) y_i + \pi_{ij}^2] \qquad (2.3)$$

If $i_0 \in \{1,\ldots,r\}$ satisfies

$$\sum_{j=1}^{u_{i_0}} (\pi_{i_0 j}-\pi_{i_0 j}^2) \geq \sum_{j=1}^{u_i} (\pi_{ij}-\pi_{ij}^2), \quad i=1,\ldots,r,$$

then the set of y_i's defined by $y_{i_0}=1$, $y_i=0$, $i \neq i_0$, maximizes (2.3) and the maximum value of (2.3) is

$$\sum_{j=1}^{u_{i_0}} (\pi_{i_0 j}-\pi_{i_0 j}^2) + \sum_{i=1}^{r}\sum_{j=1}^{u_i} \pi_{ij}^2 \tag{2.4}$$

Furthermore, with this set of y_i's, when the ij-th PSU is in the new sample, (c) cannot occur for this i,j. This is because for $i = i_0$,

$$\pi_{i_0 j} - \min\{\pi_{i_0}, y_{i_0} p_{i_0 j}\} = 0, \qquad j=1,\ldots,u_i,$$

and hence neither (b) nor (c) can occur, while for $i \neq i_0$, $y_i = 0$, and hence neither (a) nor (c) can occur. Thus, (2.2) holds and (2.4) is an exact expression for the overlap probability.

REFERENCE

Ernst, Lawrence R. (1986), "Maximizing the Overlap Between Surveys When Information Is Incomplete," European Journal of Operational Research, 27, 192-200.