VARIANCE ESTIMATION FOR THE MULTIPLICITY ESTIMATOR IN THE SERVICE BASED ENUMERATION PROGRAM

Roger Shores, Patrick J. Cantwell, and Felipe Kohn, U.S. Bureau of the Census
Roger Shores, U.S. Bureau of the Census, Washington, D.C. 20233

Key words: shelter population, soup kitchens, Bernoulli model.

1 Introduction

Service Based Enumeration (SBE) is the statistical program that the Census Bureau uses to estimate the population of persons without usual residence who use services. The methodology selected to measure this population is a multiplicity estimate of the number of times they use service facilities. This paper first presents the justification of the estimator and a derivation of its variance. The estimator of this variance then follows in a straightforward fashion. We examine the behavior of the multiplicity estimator and its variance. An important specific case is the one in which usage is assumed to follow a Bernoulli distribution. Results are presented that show what happens to the variance when the probability parameter for the Bernoulli distribution is varied.

2 SBE Methodology

2.1 The SBE Estimator

Multiplicity estimation is the methodology selected for use in the SBE program. Part of the population is enumerated on a specified day and asked about their use of services during a recent reference period. This information allows us to estimate the size of the total population using services. There are many multiplicity estimators, each relying on a different multiplicity rule. The SBE estimator relies on two usage questions to obtain the data. One question asks about shelter usage, while the other, directed at people who do not use shelters, asks about usage in soup kitchens and mobile food vans.

This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. This report is released to inform interested parties of current research and to encourage discussion.

One day is selected, and everyone is counted on that day. A multiplicity estimator for the shelter component of the estimator is given by

$$\hat{Y}_1 = \sum_{k=1}^{n} \frac{7}{A_k}, \qquad (1)$$

where n represents the number of persons enumerated at a shelter on the selected day, and A_k represents the number of days person k used a shelter during the shelter reference week, that is, the current day and the six prior days.

Enumeration at soup kitchens and mobile food vans took place the day after enumeration at shelters. The soup kitchen and mobile food van component is given by

$$\hat{Y}_2 = \sum_{h=1}^{m} \frac{7}{B_h}, \qquad (2)$$

where m represents the number of persons enumerated the next day at a soup kitchen or a mobile food van during the soup kitchen reference week, which consisted of the next day and the six prior days. B_h represents the number of days person h received a meal from a soup kitchen or mobile food van. This summation does not include people who used a shelter during the shelter reference week. The combined estimator is then

$$\hat{Y} = \sum_{k=1}^{n} \frac{7}{A_k} + \sum_{h=1}^{m} \frac{7}{B_h}. \qquad (3)$$

For more information on this estimator see Kohn and Griffin (1999).

2.2 Justifying the Estimator Statistically

In order to justify our use of the estimator in (3), we describe circumstances under which it is appropriate. In this section and the next two, we examine the properties of the shelter-only estimator, as given in (1). The results will then be extended to the combined estimator for the shelter and soup kitchen populations in section 2.5. To start we make the following assumptions:

(a) The entire population of shelter users can be

divided into eight mutually exclusive groups G_0, G_1, ..., G_7, where G_i includes all those who used shelters i times in the reference week.

(b) Each person in G_i used the shelter on i days randomly selected during the reference week.

(c) Users in the population visit shelters independently of each other.

(d) There is no response error. That is, the number of days given as the frequency of shelter use during the reference week is the true number.

Obviously, these assumptions do not hold in reality. For example, consider (b). It is likely that shelters experience heavier usage certain days of the week or different times of the month. Indeed, weather may be an important factor. Assumption (c) ignores the clustering effect of companions and of mothers with children. The most questionable assumption is (d). It is likely that many users will simply not recall how many times they have visited shelters over a week's time. However, there are few inferences we can make without these or other such assumptions.

It is worth mentioning what these assumptions do *not* imply. (1) They do *not* assume that each person in the population behaves the same way with respect to the use of shelters. That is, the probability that a person falls in G_i can vary from person to person. (2) For any individual, the mechanism for determining whether a visit is made to a shelter need not be independent over the days of the week.

Note that this population does *not* include people who *never* use shelters. The goal of SBE is not to estimate all homeless people, but simply those who use shelters (and, in section 2.5, soup kitchens). The set of people in G_0 are those who sometimes use shelters but did not use a shelter during the reference week. Among shelter users, only those in G_0 can be included in the soup kitchen and mobile food van estimate. That estimate therefore covers the people in G_0, and those who never use shelters but sometimes use soup kitchens and mobile food vans.

Let N_i be the number of people in group G_i, and let N be the size of the shelter population, that is,

$$N = \sum_{i=0}^{7} N_i .$$

From equation (1) one sees that the shelter-only multiplicity estimator is the sum of 7 over A_k, where A_k is the number of shelter visits made during the reference week for the kth person, over the population enumerated at shelters. By grouping together the A_k's corresponding to people with the same number of visits (that is, into their groups G_i), we can rewrite the estimator as $\sum_{i=1}^{7} n_i(\frac{7}{i})$, where, for i = 1, 2, ...7, n_i is the number of people *counted in the SBE* operation (out of N_i people in G_i) who visited shelters i times in the past week.

It is easy to determine the conditional distribution of the n_i. Under the assumptions above, given N_i, n_i follows a binomial distribution with N_i trials and probability of success equal to i/7. (One observes that n_7 is equal to N_7 with certainty).

Since $n_i(\frac{7}{i})$ is an obvious estimator for N_i, the shelter-only multiplicity estimator, $\sum_{i=1}^{7} n_i(\frac{7}{i})$, can be defined as $\hat{N}$. We see an immediate problem: with the shelter-only estimator, $\hat{N}$, we have not included a component to estimate N_0, the number of people in G_0. We leave the derivation of more complex methods to estimate N_0 to another paper, but present a partial remedy in section 2.5, where we investigate the combined estimator for shelters and soup kitchens.

2.3 Conditional Mean and Variance of $\hat{N}$

Because each n_i follows a binomial distribution with parameters N_i and i/7, the derivation of the expected value and variance of the multiplicity conditional estimator is straightforward:

$$E(n_i \mid \{N_0, N_1, ..., N_7\}) = N_i (i / 7) , \text{ and}$$

$$E(\hat{N} \mid \{N_0, N_1, ..., N_7\}) = \sum_{i=1}^{7} N_i = N - N_0 .$$

Clearly, $\hat{N}$ is biased downwards by the amount N_0.

Under assumptions (a) and (c), and conditional on the set $\{N_0, N_1, ..., N_7\}$, the random variables $n_1, n_2, ..., n_7$ are stochastically independent. It then follows that

$$\begin{aligned} \mathrm{Var}(\hat{N} \mid \{N_0, N_1, ..., N_7\}) &= \sum_{i=1}^{7} \mathrm{Var}(n_i \mid \{N_0, N_1, ..., N_7\}) (7/i)^2 \\ &= \sum_{i=1}^{7} N_i (i/7) (1 - i/7) (7/i)^2 \\ &= \sum_{i=1}^{7} N_i (7-i) / i \end{aligned} \qquad (4)$$

We obtain a straightforward estimator for this variance by estimating the population components N_i :

$$\hat{\mathrm{Var}}\,(\hat{N} \mid \{N_0, N_1, ..., N_7\}) = \sum_{i=1}^{7} n_i\,(7/i)\,(7\text{-}i) / i$$

$$= \sum_{i=1}^{7} n_i\,7\,(7\text{-}i) / i^2 . \qquad (5)$$

Finally, we can obtain the conditional variance of this variance estimator:

$$\mathrm{Var}\,(\hat{\mathrm{Var}}\,(\hat{N} \mid \{N_0, N_1, ..., N_7\} \mid \{N_0, N_1, ..., N_7\})) = \sum_{i=1}^{7} N_i\,(i/7)\,(1 - i/7)\,[7\,(7\text{-}i) / i^2\,]^2$$

$$= \sum_{i=1}^{7} N_i\,(7\text{-}i)^3 / i^3 . \qquad (6)$$

As one would expect, for a fixed total number of people $N - N_0$, the true variance increases as the people make fewer visits to shelters during the reference week. In that case, fewer shelter people tend to be enumerated, and the weights applied to their records, 7/i, tend to be larger.

A simple estimator for the variance of $\hat{\mathrm{Var}}\,(\hat{N} \mid \{N_0, N_1, ..., N_7\})$ in (6) can be given by inserting estimates, $n_i\,(7/i)$, for each N_i .

2.4 Unconditional mean and variance

The results in the previous section demonstrate what happens conditional on the population sizes $N_1, N_2, ..., N_7$. But the behavior of the estimator $\hat{N}$ also depends on the stochastic mechanism that produces the values of the N_i's. After all, someone who visits a shelter four times one week may well visit twice in another week. How does this variability affect the distribution of $\hat{N}$?

The unconditional mean of $\hat{N}$ is as follows:

$$E\,(\hat{N}) = E\,[\,E\,(\sum_{i=1}^{7} n_i\,(7/i) \mid \{N_0, N_1, ..., N_7\})\,,$$

where the outside expectation is taken over the distribution of possible values of the vector $\{N_0, N_1, ..., N_7\}$. As the terms $n_i\,(7/i)$ are conditionally unbiased for the N_i, we can write

$$E\,(\hat{N}) = E\,[\sum_{i=1}^{7} N_i\,] = E\,[\,N - N_0\,] = N - E\,[\,N_0\,].$$

Thus the mean of $\hat{N}$ has a downward bias equal to the expected value of N_0, the expected size of the shelter population who do not make a visit in the reference week (G_0). The unconditional variance can be derived similarly:

$$\mathrm{Var}\,(\hat{N}) = \mathrm{Var}\,[\,E\,(\sum_{i=1}^{7} n_i\,(7/i) \mid \{N_0, N_1, ..., N_7\})\,] + E\,[\,\mathrm{Var}\,(\sum_{i=1}^{7} n_i\,(7/i) \mid \{N_0, N_1, ..., N_7\})\,]$$

$$= \mathrm{Var}\,[\sum_{i=1}^{7} N_i\,] + E\,[\sum_{i=1}^{7} N_i\,(7\text{-}i) / i\,]$$

$$= \mathrm{Var}\,[\,N - N_0\,] + \sum_{i=1}^{7} E(N_i)\,(7\text{-}i) / i$$

$$= \mathrm{Var}\,[\,N_0\,] + \sum_{i=1}^{7} E(N_i)\,(7\text{-}i) / i. \qquad (7)$$

The leading component of the variance in (7), $\mathrm{Var}\,[N_0\,]$, may be small if the term N_0 tends to be small. Note that the result in (7) requires only the assumptions made in section 2.2. It is not necessary that different people in the population visit shelters in the same way or with the same probabilities.

2.5 Extending to the combined shelter and soup kitchen estimator

To this point we have investigated the distribution of the shelter-only estimator. The extension to the combined shelter and soup kitchen multiplicity estimator is important because we can account for (i) people who visit shelters at times, but did not during the reference week, and (ii) people who never frequent shelters, but who sometimes visit soup kitchens.

Let us recall the procedure applied *at soup kitchens:* all people enumerated there are asked two questions: (1) how many times in the reference week they visited a soup kitchen, and (2) whether they visited a shelter at any time during the reference week. If the answer to (2) is "yes," their response does not contribute to the soup-kitchen component of the combined estimator, as they were *represented in the shelter population* already--whether or not they were enumerated at a shelter. The idea is to represent each person, that is, give him or her a chance to be enumerated, *but only once.* If the answer to (2) is "no," they are assigned a weight equal to seven times the reciprocal of their response to (1), similar to the weighting used in the shelter-only estimator. The result is the combined estimator in equation (3).

To justify the combined estimator, we first define the population more generally than before. We include all people who sometimes use a shelter *or* a soup kitchen *or both*, although they may have used neither in the reference week leading to enumeration day. This population can be divided into 64 groups G_{ij} of size N_{ij},

where G_{ij} includes all those who visited a shelter i times during the reference week, and also visited a soup kitchen j times during its week. In the groups G_{0j}, j = 0, 1, ..., 7, are those people who did not use a shelter during the reference week, including those who sometimes use a shelter and those who never do. According to our definition of the population, those in G_{00} sometimes use a shelter or a soup kitchen.

Because we exclude from the soup-kitchen component of the estimator any census respondents who used a shelter in the reference week, we can limit ourselves to estimating the following components of N: $N_1, N_2, ..., N_7$, $N_{01}, N_{02}, ...,$ and N_{07}. All persons in G_{ij}, where $i > 0$, *do not* contribute toward the second summation in (3), the soup-kitchen component, *but are represented* in the estimation of N_i through the first component--whether or not they were enumerated at a shelter.

We extend the assumptions given above in section 2.2 analogously to the use of soup kitchens. Then the combined estimator can be written as

$$\hat{N} = \sum_{i=1}^{7} n_i (7/i) + \sum_{j=1}^{7} n_{0j} (7/j),$$

where the n_{0j} are the number of people in G_{0j} enumerated in the SBE operation at soup kitchens. The results derived in the previous sections carry forward analogously. Conditional on the set of population sizes $\{N_1, N_2, ..., N_7; N_{01}, N_{02}, ..., N_{07}\}$,

$$E(\hat{N} \mid \{N_1, N_2, ..., N_7; N_{01}, N_{02}, ..., N_{07}\}) = \sum_{i=1}^{7} N_i + \sum_{j=1}^{7} N_{0j} = N - N_{00}, \text{ and}$$

$$\mathrm{Var}(\hat{N} \mid \{N_1, N_2, ..., N_7; N_{01}, N_{02}, ..., N_{07}\}) = \sum_{i=1}^{7} N_i (7-i)/i + \sum_{j=1}^{7} N_{0j} (7-j)/j.$$

This suggests the following conditional estimator of the variance of N hat:

$$\hat{Var}(\hat{N} \mid \{N_1, N_2, ..., N_7; N_{01}, N_{02}, ..., N_{07}\}) = \sum_{i=1}^{7} n_i\, 7\,(7-i)/i^2 + \sum_{j=1}^{7} n_{0j}\, 7\,(7-j)/j^2. \quad (8)$$

As with the shelter-only estimator, the conditional variance of the variance estimator is easily obtained:

$$\mathrm{Var}(\hat{Var}(\hat{N}) \mid \{N_1, N_2, ..., N_7; N_{01}, N_{02}, ..., N_{07}\}) = \sum_{i=1}^{7} N_i (7-i)^3/i^3 + \sum_{j=1}^{7} N_j (7-j)^3/j^3 \quad (9)$$

Unconditionally, the mean and variance of $\hat{N}$ are

$$E(\hat{N}) = E[\sum_{i=1}^{7} N_i + \sum_{j=1}^{7} N_{0j}] = N - E[N_{00}],$$

and

$$\mathrm{Var}(\hat{N}) = \mathrm{Var}[N_{00}] + \sum_{i=1}^{7} E(N_i)(7-i)/i + \sum_{j=1}^{7} E(N_{0j})(7-j)/j. \quad (10)$$

Again we see that the estimator is biased downward, now by $E[N_{00}]$, the mean number of people who frequent shelters or soup kitchens, but who visit *neither* during the appropriate reference week. This number is generally smaller than $E[N_0]$, the bias in the shelter-only estimator. If we are willing to apply information about the behavior of people in the target population, we can model the distribution of N_{00} and the set $\{N_1, N_2, ..., N_7; N_{01}, N_{02}, ..., N_{07}\}$, and thereby predict the unconditional performance of the combined estimator, $\hat{N}$. In fact, under reasonable assumptions the unconditional mean and variance of N_{00} may contribute only a very small part of the total mean and variance.

3 Constant Visit Probabilities Over the Population

To investigate the behavior of the multiplicity estimator under actual conditions, we add one assumption to those made in section 2.2:

(e) The probability, f_{ij}, i = 0, 1, ..., 7, that a person makes i visits to a shelter and j visits to a soup kitchen during the respective reference weeks, is the same for each person.

Note that we are not yet making any further claims about the day-to-day behavior of the individuals beyond what has already been assumed in section 2.2.

Consider first the shelter-only estimator. Analogous to previous notation, for i = 0, 1, ..., 7, let f_i be the sum of the f_{ij}'s as j runs from 0 to 7. Under (e) and the prior assumption of independent shelter use ((c) in 2.2), for a population of N people who sometimes use shelters, the set $\{N_0, N_1, ..., N_7\}$ follows a multinomial distribution with N trials and probability parameters $f_0, f_1, ..., f_7$. Thus each N_i has a binomial distribution with parameters N and f_i. The conditional results of section 2.3 pertaining to the multiplicity estimator continue to hold. But now we can derive its unconditional mean and variance as well. Substituting into the equations in section 2.4, we conclude that

$$E(\hat{N}) = (N - N \cdot f_0) = N(1 - f_0), \text{ and}$$

$$\mathrm{Var}(\hat{N}) = N f_0 (1 - f_0) + \sum_{i=1}^{7} N f_i (7-i)/i.$$

If f_0 is small, the relative bias is small, and the first component of the unconditional variance is small relative to the other components.

Extending these results to soup-kitchen enumeration is straightforward. The number in the population who make no visits to shelters or soup kitchens during the reference weeks, N_{00}, follows the binomial law with parameters N and f_{00}. The bias in the combined estimator is $E[N_{00}] = N \cdot f_{00}$, and the variance is

$$\mathrm{Var}(\hat{N}) = N f_{00}(1 - f_{00}) + \sum_{i=1}^{7} N_i f (7-i)/i + \sum_{j=1}^{7} N f_{0j} (7-j)/j. \quad (11)$$

4 The Bernoulli Model for Visit Behavior

In this section we discuss the case where individual service usage follows a Bernoulli distribution. We make one last assumption:

(f) On any day in the reference week, each person is assumed to use a shelter (soup kitchen) with probability p_1 (p_2), with behavior independent from day to day and over facilities.

Thus for any person the number of visits to a shelter (soup kitchen) over the seven days is binomially distributed with parameters 7 and p_1 (p_2); the f_i's and f_{0j}'s above are replaced by values of the binomial probability function. It follows easily that

$$E(\hat{N}) = N(1 - f_{00}) = N[1 - (1-p_1)^7(1-p_2)^7],$$

and from equation (11),

$$\mathrm{Var}(\hat{N}) =$$

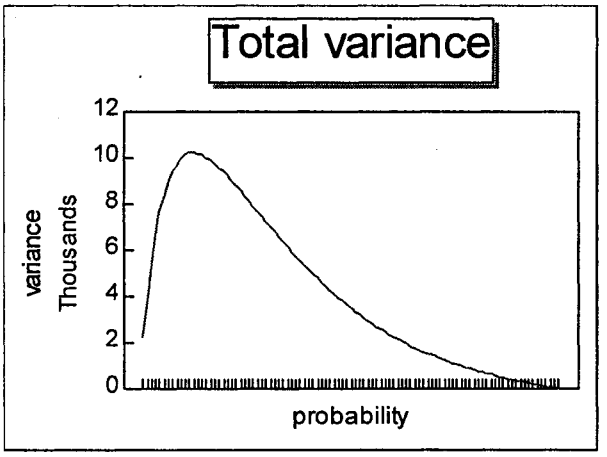
$$N[(1-p_1)^7(1-p_2)^7][1 - (1-p_1)^7(1-p_2)^7] \quad (12.1)$$

$$+ \sum_{i=1}^{7} N \binom{7}{i} p_1^i (1-p_1)^{7-i} (7-i)/i \quad (12.2)$$

$$+ \sum_{i=1}^{7} N (1-p_1)^7 \binom{7}{j} p_2^j (1-p_2)^{7-j} (7-j)/j. \quad (12.3)$$

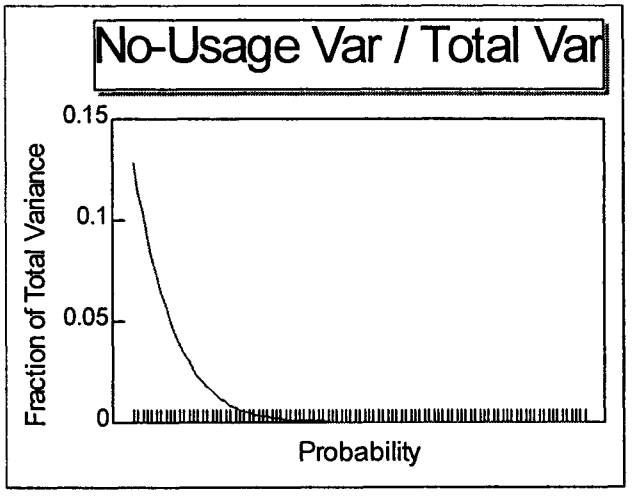
We will make the simplifying assumption that usage at shelters and soup kitchens occurs with the same probability; that is, $p_1 = p_2 = p$. Different probabilities of service usage can produce very different estimates of the unconditional variance and of related statistics. These statistics are proportional to N, but it is illustrative to hold N constant while varying p. The following graphs give the variance and other statistics as a function of p, while holding N constant at 2,500.

The first graph gives the total unconditional variance; it reaches a maximum of about 10,200 at p = 0.13. As p is reduced from 1, the variance at first increases because the counts in the lower usage categories for the shelter and soup kitchen variances will increase, and, as we have seen, the lower usage categories have more of an effect



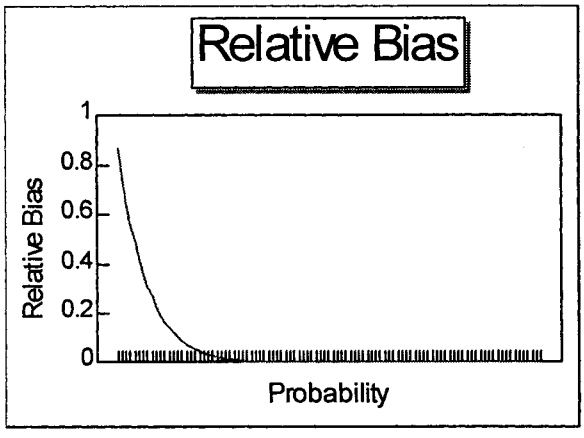
on the variance than do the higher usage categories. Eventually, however, as p is reduced further, so many people fall into the no-usage category that the variance of the other two components begins to decrease, so that the total variance will decrease.

The no-usage variance also decreases once p becomes small enough. It does, however, assume a larger fraction of the total as p becomes smaller. The next graph presents that variance as a fraction of the total variance.



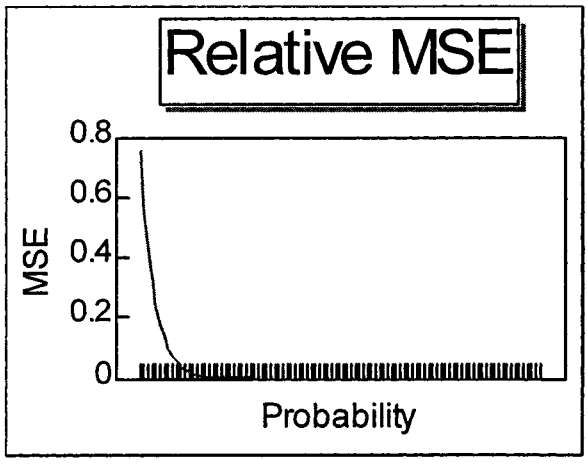
It shows that the no-usage component of the variance is only significant when p is small. At p = 0.13, it is only about 3% of the total, and at p = 0.20, it has already fallen to approximately 1%.

The next graph shows the relative bias of the estimator, $E(N_{00}) / 2500 = (1 - p)^{14}$.

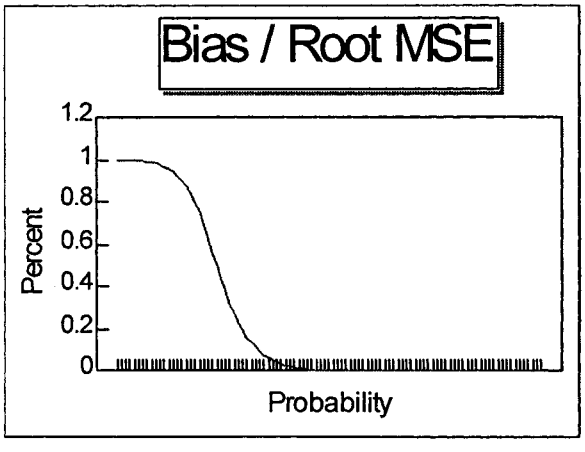


It too quickly falls to insignificance as p increases, equaling, for example, approximately 0.02 at p = 0.25.

Next, we see a graph of the relative mean square error.



Finally, we see the bias as a fraction of the square root of the MSE.



The relative bias is initially substantial. At p = 0.01, for example, it equals 87%. As one can see from the prior graph, it falls quickly; as long as p is equal to at least 0.3, it will be under 1%. The relative mean square error declines in a similar manner. Thus the bias is, at the lowest usage probabilities, a substantial fraction of the root MSE, making up essentially all of it for probabilities of about 0.1 or less. Once p has risen to 0.3, however, this fraction has fallen to approximately 23% and is declining rapidly as a proportion of the total. By p = 0.4 it contributes only about 3%, and at p = 0.5 it has become negligible as a fraction of the root MSE.

To make this analysis more realistic, we can let p_1 differ from p_2. The general results will be similar, but we can examine the effect of differences in usage at shelters and soup kitchens. Such results are not included here because of limited space.

5 Continuing Research

Currently we are pursuing several areas of research:

- Estimating the "no-usage" component of N, that is, N_0 or N_{00}, from the enumeration data.
- Modeling the behavior of the individuals in the population allowing for different probabilities of service usage from person to person.
- Determining the distribution of $\{N_0, N_1, ..., N_7\}$ under various behavior models.
- Approximating the actual distribution of $\{N_0, N_1, ..., N_7\}$ by similar distributions whose parameters can be reasonably estimated.
- Evaluating the statistical properties of the derived estimators, that is, $\hat{N}$ and $\widehat{\mathrm{Var}}(\hat{N})$, by using simulations.

References

Kohn, F. and Griffin, R., (1999). "Multiplicity Estimators Applied in the Service Based Enumeration Program," *Proceedings of the Section on Survey Research Methods, American Statistical Association (to appear).*