SAMPLING AND ESTIMATION FOR THE HOMELESS POPULATION

Eric Schindler, Richard Griffin, and Alfredo Navarro[1]
Eric Schindler, Bureau of the Census, Washington, DC 20233

KEY WORDS: Estimation, Response Bias, Capture/Recapture

The Census Bureau is conducting research into methodologies for estimating the size of the homeless population. These alternative statistical methods concentrate on shelters, soup kitchens, and other selected locations. Two classes of estimates are being considered. One estimate, based on capture/recapture methods, matches results from samples for two or more days to produce dual-system estimates (DSEs). The second type of estimate avoids matching, but relies on respondents' answers to "site use history" questions. Both methods are consistent with the Census 2000 research goal of studying sampling and statistical methods to "count" the population.

Unlike programs designed specifically to estimate the number and characteristics of the homeless, the Decennial Census must try to enumerate and assign a geographic location to 250 million persons. Census procedures for the homeless can be directed only at persons who will not be counted elsewhere. While the most efficient procedures for estimating the homeless population may rely entirely on sampling, it is necessary to attempt a complete count at least once in order to give each person an opportunity to be enumerated. Additional procedures can make use of sampling techniques to get the most reliable estimates for the available resources.

STRATEGIES WITH MATCHING

The Post-Enumeration Survey (PES) of the 1990 Census used capture/recapture techniques to estimate the undercount (See Wolter, 1986). In a first round subjects are counted by complete enumeration or sampling. In a later sample or samples, persons are matched to their original enumeration. A DSE is used to estimate the total population. If 1000 homeless persons in a city are enumerated on Day 1, and 200 of the 500 persons enumerated on Day 2 are matched to Day 1 reports, the statistical inference is that the 1000 persons counted on Day 1 are 40% of the homeless population. This is equivalent to the capture/recapture models, as described in Seber (1982), which are used to estimate wildlife numbers and densities. A similar usable method plants a known number of persons to act as "homeless" persons during a single enumeration. The proportion of the planted persons found provides the basis for a DSE. (See Laska and Meisner.)

In order to develop a setting for the DSE, assume that 50% of the homeless population can be enumerated at shelters, soup kitchens, and other locations providing services on a previously developed list on a randomly selected day. Assume also that the persons found can be identified as homeless, and enumerated so as to be identifiable if sampled again. This means that persons counted at more than once on any given day can be unduplicated.

Consider a series of s samples with sizes n_1, n_2, ... , n_s from a total homeless population of size N. Let m_i be the number of persons in day i's sample who have already been enumerated. Then $u_i = n_i - m_i$ is the number of not previously enumerated persons found on day i. Define $M_i = \sum_{j=1}^{i-1} u_j$, for $i = 1,2,\ldots,s+1$ to be the number of enumerated individuals just before the ith sample is selected. Note that $m_1 = M_1 = 0$ and $u_1 = M_2 = n_1$. Also, M_{s+1} is the number of enumerated individuals at the end

[1] Eric Schindler and Alfredo Navarro are Mathematical Statisticians in the Decennial Statistical Studies Division of the Census Bureau. Richard Griffin is now Chief Methodologist at Chilton Research Services. This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau.

of the s days. Let p=0.5 be the assumed enumeration probability for any individual on any given day. Let q=1-p. Then $n_i \sim B(N,p)$, so $E(n_i)=Np$.

Define:

$$\beta_{ij} = \{ \begin{array}{ll} 1 & \textit{if person j is first enumerated} \\ & \textit{just before day i, i.e. on day i-1} \\ 0 & \textit{otherwise} \end{array}$$

Then $M_i = \sum_{j=1}^{N} \beta_{ij}$ and, therefore,

$$\begin{aligned} E(M_i) &= \sum_{j=1}^{N} E(\beta_{ij}) = \sum_{j=1}^{N} P(\beta_{ij}=1) \\ &= \sum_{j=1}^{N} (1-q^{i-1}) = N(1-q^{i-1}) \end{aligned} \qquad (1)$$

Let $\hat{N}_s$ be the maximum likelihood estimate of the total homeless population if s days are sampled. When s=2, Seber gives the "Peterson estimate," equivalent to the DSE, $\hat{N}_2 = \frac{n_1\, n_2}{m_2}$.

Seber also gives the following approximation for the coefficient of variation (CV) of $\hat{N}_2$:

$$CV(\hat{N}_2) = \sqrt{\frac{N}{n_2\, M_2}} = \sqrt{\frac{N}{n_2\, n_1}}$$

Each additional day of sampling provides a new DSE with all the previous enumeration collapsed to form the first stage. The DSE $\hat{N}_i$ can be formed after each day has:

$$CV(\hat{N}_i) = \sqrt{\frac{N}{n_i\, M_i}} \quad \text{or} \quad V(\hat{N}_j) = \frac{N^3}{n_i\, M_i} \ .$$

A weighted average of the $\hat{N}_i$ values with weights inversely proportional to the variance of the $\hat{N}_i$ estimates produces an estimate with the same variance as the maximum likelihood estimate derived by an iterative method described in Seber.

$$\text{Define} \quad \hat{N}_i^* = \sum_{j=2}^{i} \frac{\frac{1}{V(\hat{N}_j)}}{\sum_{j=2}^{i} \frac{1}{V(\hat{N}_j)}} \hat{N}_j \ .$$

Ignoring the covariance we derive:

$$V(\hat{N}_i^*) = \frac{N^3}{\sum_{j=2}^{i} n_j\, M_j} \quad \text{or} \quad CV(\hat{N}_i^*) = \sqrt{\frac{N}{\sum_{j=2}^{i} n_j\, M_j}}$$

This is the CV of the maximum likelihood estimate of the total homeless population after i days of enumeration of sites. Using expected values, we derive:

$$CV(\hat{N}_i^*) \cong \sqrt{\frac{1}{Np\sum_{j=2}^{i} (1-q^{j-1})}} \qquad (2)$$

Assuming a uniform 50% enumeration rate, Table 1, using equations (1) and (2) shows the number of persons who would be found after each of the first 4 days for three population sizes corresponding to the homeless populations we might find in small, medium, and large cities. The coefficients of variation are also given.

Table 1: Expected number of persons found after 1, 2, 3, or 4 days with a uniform 50% enumeration rate.

True Pop	Found After Day			
	1	2	3	4
600 CV	300	450 .082	525 .061	563 .040
6000 CV	3000	4500 .026	5250 .019	5625 .013
60000 CV	30000	45000 .008	52500 .006	56250 .004
%	50%	75%	88%	94%

The homeless population does not have a uniform 50% enumeration rate. There appears to be a consensus in the literature that about half of the homeless persons who use the facilities in an area can be found by enumerating shelters and soup kitchens on any given day. If enumeration continues for a week, about 70% can be found. The remaining 30% of the homeless population is more difficult to find. Estimates of the proportion of those homeless who use facilities at least some of the time which can be found by enumerating such locations every day for a month vary from

85% to 95%. An additional portion of the homeless can be found at semi-permanent "encampments". The last 1% to 10% have little contact with shelters or soup kitchens and would be difficult to enumerate with the procedures available to the Decennial Census.

For illustrative purposes, two test populations are assessed in the following examples. In order to have even a small percentage of persons who use facilities still missed after a month, a substantial percentage can use facilities at most a few times a month. If the proportion found on a given day is p, then after n days $1-(1-p)^n$ will still be uncounted. If p=.25 and n=28, only 0.03% will not be found in a month. If p=.10 and n=28, 5% will not be found in a month.

Population I consists half of persons who can be found 92% of the time by enumerating both shelters and soup kitchens on a single day. The other half can be found at service sites only 8% of the time. For this population, 50% can be found in a day, 72% in a week, and 95% in a month. Population II consists half of persons using facilities 90% of the days of the month, one-sixth of persons using facilities 24% of the days, and one-third of persons using facilities only 3% of the days, or about once a month. For this population, 50% can be found in a day, 71% in a week, and 86% in a month.

As shown in Table 2, Test Populations I and II have lower overall enumeration rates after the first day than the uniform population shown in Table 1. The CVs of the maximum likelihood estimates derived from (2) for these cells are also given. The estimation of Cvs for the total population requires independence between the estimation cells or strata. Estimating the number of homeless in each stratum requires classifying people into the two or three strata based on frequency of facilities usage, probably with a question such as: "How often do you use facilities for the homeless: Almost everyday, several times a week, several times a month, or rarely?" As the population is increased from 6,000 to 60,000, the Cvs decrease by a factor of about 3.

Table 2: Expected number of persons found and coefficients of variation after 1, 2, 3, or 4 days with non-uniform enumeration rates.

Pr	Pop	Found After Day			
		1	2	3	4
PPOPULATION I					
.92	3000	2760	2981	2999	3000
	CV	--	.02	.01	.01
.08	3000	240	461	664	851
	CV	--	.23	.13	.10
Tot	6000	3000	3442	3663	3851
	%	50%	57%	61%	64%
	CV	--	.11	.07	.05
POPULATION II					
.90	3000	2700	2970	2997	3000
	CV	--	.02	.01	.01
.24	1000	240	422	561	666
	CV	--	.13	.10	.08
.03	2000	60	118	175	229
	CV	--	.75	.43	.31
Tot	6000	3000	3510	3733	3895
	%	50%	59%	62%	65%
	CV	--	.25	.14	.10

For Population I, with half the persons locatable 8% and half 92% of the time, it appears that reasonable estimates of the total homeless population can be obtained by selecting for enumeration two days at random.

Wolter (1986) provides a coverage error model applicable to this situation. This model produces a maximum likelihood estimate for stratified DSEs. Let N_h be the homeless population and p_h be the enumeration probability in stratum h. On day 1, all sites are enumerated finding, say $\hat{n}_{h1}$ persons in stratum h. Then $E(\hat{n}_{h1})=N_h\,p_h$. On day 2 a sampling fraction f is used to sample sites and homeless persons in the selected sites are enumerated. Then let $\hat{n}_{h2}$ and $\hat{n}_{h12}$ be estimates from the one stage simple random cluster sample of sites of the number of homeless persons and the number of matches in stratum h. The DSE is:

$$D\hat{S}E_h = \frac{\hat{n}_{h1}\,\hat{n}_{h2}}{\hat{n}_{h12}}$$

Assuming autonomous independence, Wolter approximates the variance of $D\hat{S}E_h$,including both model and sampling error, by:

$$VAR(D\hat{S}E_h) = \frac{N_h(1-p_h)}{p_h^2}\ \frac{1-p_h f}{f} \qquad (3)$$

Table 3 uses (3) to give the coefficient of variation for the DSE for each stratum for both a 100% enumeration and 50% sampling on day two. If there is little covariance between the strata defined by the simple question on the history of site usage, the CV of the total population can also be easily derived. Also, given is the CV if all persons had a 50% enumeration probability. The results for Population I are much better than those for Population II.

Table 3: Coefficients of variation for DSEs with 2 population levels, 2 day two sampling rates, and 2 enumeration probability distributions.

Prob	Pop	CV100%	CV50%
Uniform 50% Enumeration Rate			
.50	6000	.0129	.0233
.50	60000	.0040	.0070
POPULATION I			
.92	3000	.0015	.0058
.08	3000	.2099	.3033
Total	6000	.1049	.1516
.92	30000	.0005	.0018
.08	30000	.0663	.0959
Total	60000	.0331	.0479
POPULATION II			
.90	3000	.0020	.0067
.24	1000	.1001	.1523
.03	2000	.7229	1.0303
Total	6000	.2415	.3444
.90	30000	.0006	.0021
.24	10000	.0316	.0481
.03	20000	.2286	.3258
Total	60000	.0763	.1089

Matching the enumerated homeless persons over several days and sites will result in nonsampling errors. Enumeration on one day may alter behavior on later days, especially for persons enumerated several times, negating independence assumptions. If these problems can be controlled and if the actual homeless population is more like Population I (i.e., not too many rare service users), the DSE with cluster sampling of sites on the second day can produce reliable estimates of the total homeless population. Since there will be substantial nonsampling error, it does not make sense to set the Day 2 sampling rate to achieve expected coefficients of variation lower than about 10%.

STRATEGIES WITHOUT MATCHING

It may be impossible to reduce the bias from matching and other errors with capture/recapture or dual-system estimation methods to acceptable levels. This section describes two enumeration methodologies which based on responses to "site use" questions. Two different questions lead to two equivalent formulations. Unbiased estimates require correct responses to these questions. More precise answers are needed than for the stratification question used for the matching strategy. Even small response biases can lead to large underestimates of the population.

QUESTION: **How many days in a month do you use service sites (soup kitchens, shelters, clinics, etc) for the homeless in this area?**

Let H=28 be the number of days in a month and N the number of sites in the area. There is a population of HN site days from which to select a sample. Let M_{ij} be the number of persons enumerated at site j on day i. Let A_{ijk} be the number of days person k enumerated at site j on day i reports using sites in the area. To avoid duplications for persons who use more than one site per day, A_{ijk} can be set to the total number of sites visited in a month and persons can be asked how many sites they use on an "average" day. Obtaining the data for this revised question will add response bias, but it is necessary to avoid overestimates.

$Y_{ij} = \sum_{k=1}^{M_{ij}} \frac{1}{A_{ijk}}$ is the value of site j on day i counting each person inversely proportionally to the number of sites used per month. If a complete count were taken each day, all persons using any site during the month would be counted exactly once. Now consider a two stage cluster sample consisting of a simple random sample of h days and, for each day, a simple random sample of n sites. The estimator,

$$\hat{Y} = \frac{HN}{hn}\sum_{i=1}^{h}\sum_{j=1}^{n} Y_{ij} \ ,$$

has the appropriate expected value,

$$E\hat{Y} = Y = \sum_{i=1}^{H}\sum_{j=1}^{N}\sum_{k=1}^{M_{ij}} \frac{1}{A_{ijk}} \ , \text{ and}$$

$$Var(\hat{Y}) = (HN)^2[(1-\frac{h}{H})\frac{S_b^2}{h} + (1-\frac{n}{N})\frac{\bar{S}_w^2}{hn}]$$

where:

$$S_b^2 = \frac{1}{H-1}\sum_{i=1}^{H}(\frac{1}{N}\sum_{j=1}^{N}Y_{ij} - \frac{1}{HN}\sum_{i=1}^{H}\sum_{j=1}^{N}Y_{ij})^2$$

is the variance among day means, and

$$\bar{S}_w^2 = \frac{1}{H}\sum_{i=1}^{H}(\frac{1}{N-1}\sum_{j=1}^{N}(Y_{ij}-\frac{1}{N}\sum_{j=1}^{N}Y_{ij})^2)$$

is the variance among sites within days.

It is probably not realistic to expect good estimates from the homeless of the number of times soup kitchens or shelters are used in a month. An alternative "site use" question which may be easier to answer is:

QUESTION: How many days ago was the last time you used a service site for the homeless in this area?

This question assumes that the persons in the homeless population, like persons in the general population, should be able to estimate the last time they used a facility more easily than the number of times they use facilities in a month. If so, there will be less response bias to this question than to the reformulated first question.

Define H, N, M_{ij}, h, and n as above. Let t_{ijk} be the number of days since the last time person k found in site j on day i was at a site in the area. If $t_{ijk} \geq 28$, set $t_{ijk} = 28$. Note that someone who had already been in another site on that day should answer 0.

Set $X_{ij} = \sum_{k=1}^{M_{ij}} t_{ijk}$ and define the estimator:

$$\hat{X} = \frac{N}{hn}\sum_{i=1}^{h}\sum_{j=1}^{n}X_{ij} = \frac{N}{hn}\sum_{i=1}^{h}\sum_{j=1}^{n}\sum_{k=1}^{M_{ij}}t_{ijk} \ .$$

This is called a "time-to-last capture model". The longer the time since the last potential enumeration, the higher the "weight" of the respondent. $Var(\hat{X})$ is similar to $Var(\hat{Y})$. Charles Alexander of the Census Bureau has shown that ignoring response bias this estimator is unbiased, provided (1) over repeated 28 day cycles, the number of persons using sites on day i equals the number using sites on day i+28, and (2) the number of persons last using a site k or more days ago on day i is equal to the number last using a site k or more days ago on day i+28. Utilization patterns, especially with significant weather changes around Census time, may contradict these assumptions.

Note that if we assume $t_{ijk} = 28/A_{ijk}$, then,

$$\hat{X} = \frac{N}{hn}\sum_{i=1}^{h}\sum_{j=1}^{n}\sum_{k=1}^{M_{ij}}\frac{28}{A_{ijk}} = \frac{28N}{hn}\sum_{i=1}^{h}\sum_{j=1}^{n}\sum_{k=1}^{m_{ij}}\frac{1}{A_{ijk}} = \hat{Y}$$

A major source of bias for this approach is that persons may not remember the last time they used a facility. Sudman and Bradburn (1974) discuss a simple model of the effect of time on memory. Telescoping error occurs when a respondent misreports the time of an event, primarily in the direction of remembering the event as having occurred more recently than it actually did. Telescoping would decrease estimates of the homeless population using time-to-last capture methods. According to Weber's Law in Sudman and Bradburn (1974), errors in perception of time are a function of the logarithm of the time period.

Let t_{ijkr} be the response of person k found in unit j on day i to the question concerning the number of days since the last visit to a unit. Let t_{ijkA} be the actual answer. Under telescoping, $t_{ijkr} < t_{ijkA}$. By Weber's Law the net absolute error is log bt, where b is called the telescoping parameter. The relative error in the length of the time report is r.e.$_t$ = (log bt)/t. We can model t_{ijkr} by:

$$t_{ijkr} = t_{ijkA} - e_{ijkr} \text{ where } e_{ijkr} \sim (\log bt_{ijkA}, \sigma^2) \ .$$

$$\hat{X} = \frac{N}{hn}\sum_{i=1}^{h}\sum_{j=1}^{n}\sum_{k=1}^{M_{ij}} t_{ijkr}$$

$$E(\hat{X}-X) \cong$$

$$\frac{1}{H}\sum_{i=1}^{H}\sum_{j=1}^{N}\sum_{k=1}^{M_{ij}}(t_{ijkA}-\log bt_{ijka})-\frac{1}{H}\sum_{i=1}^{H}\sum_{j=1}^{N}\sum_{k=1}^{M_{ij}}t_{ijkA}$$

$$= -\frac{1}{H}\sum_{i=1}^{H}\sum_{j=1}^{N}\sum_{k=1}^{M_{ij}}\log bt_{ijkA}$$

$$= -\frac{1}{H}\sum_{s=1}^{28}a_s\ N_s \log bs$$

where N_s is the number of persons for whom it has been s days since the last expected enumeration and a_s=28/s is the number of days each of these persons would be enumerated in a month. Note that $X \cong \frac{1}{H}\sum_{s=1}^{28} s\ a_s\ N_s$.

We have no estimate of b, but the telescoping bias can be substantial for b>1. More bias is created by persons who use facilities every two or three days and report slightly less than by persons who use facilities rarely and report several days less. The two test populations were allocated to visit units every so many days consistent with the previous assumptions. For the last two columns in Table 4, it is assumed that there is no telescoping bias for persons who last visited a unit the previous day. If reporting 1 day as 0 days occurs, the estimated values for b > 1, as shown in the first two columns, are severely biased. For the last two columns in Table 4, we assume that there is no telescoping by persons last visiting a unit the previous day.

Table 4: Expected Values of Estimates with Telescoping Bias for a Population of 6,000 for several Telescoping Parameters

b	Persons last at site yesterday may telescope report to today		Persons last at site yesterday will not telescope report to today	
	POP I	POP II	POP I	POP II
.5	5811	5858	5811	5858
1	5684	5677	5684	5677
2	4781	4774	5558	5496
5	3588	3580	5390	5257
10	2684	2677	5264	5076

CONCLUSIONS

Three methods for estimation of the homeless population have been discussed. Each will have to be evaluated for operational feasibility for the enumeration of the homeless population within the context of the decennial census. All of the methods need large amounts of time, effort, special training, and money. All are subject to unknown levels of nonsampling error and bias. All will miss a small proportion of the homeless. The difficulty of matching persons over several days of enumeration at multiple sites in an area probably may make the capture/recapture method with dual-system estimation the least appealing approach. Site use history approaches can work if response error can be controlled, perhaps more so for persons using sites frequently than for occasional or rare users.

REFERENCES

Laska, E.M. and Meisner, M, (?), Using Plants to Estimate the Size of a Population from a Single Sample, Nathan Kline Institute for Psychiatric Research

Seber, G.A.F., (1992), The Estimation of Animal Abundance, Charles Griffin and Company Limited.

Sudman, S. and Bradburn, N., (1974), Response Effects in Surveys.

Wolter, K.M., (1986), "Some Coverage Error Models for the Census Data", Journal of the American Statistical Association, Number 81, pp 338-346.