American Journal of Epidemiology
Copyright © 2007 by the Johns Hopkins Bloomberg School of Public Health
All rights reserved; printed in U.S.A.

Vol. 165, No. 11
DOI: 10.1093/aje/kwm005
Advance Access publication March 30, 2007

Practice of Epidemiology

Estimating Population Size with Two- and Three-Stage Sampling Designs

Jacqueline E. Tate[1,2] **and Michael G. Hudgens**[3]

[1] Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, NC.
[2] MEASURE Evaluation Project, Carolina Population Center, University of North Carolina at Chapel Hill, Chapel Hill, NC.
[3] Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC.

Received for publication February 1, 2006; accepted for publication December 5, 2006.

Reliable estimates of population size are important for developing and monitoring health programs in at-risk populations. Laska, Meisner, and Siegel (*Biometrics* 1988;44:461–72) developed an unbiased estimator for the size of a population at a single venue based on a single sample. Because many populations of interest are not contained within a single venue, this article generalizes the Laska, Meisner, and Siegel estimator to incorporate two- and three-stage sampling designs and enable estimation of total population size over multiple venues. Use of the estimator with two- and three-stage sampling designs is illustrated with examples that estimate the size of a population of individuals who socialize over a 4-week period at public venues where transmission of human immunodeficiency virus and other sexually transmitted infections is likely to occur.

population size; sampling studies; multistage sampling

Abbreviations: HIV, human immunodeficiency virus; LMS, Laska, Meisner, and Siegel; PLACE, Priorities for Local AIDS Control Efforts.

Obtaining size estimates of populations at risk of human immunodeficiency virus (HIV) is critical for planning, implementing, and evaluating intervention programs. Prevention of new HIV infections requires targeted contact with a large proportion of the at-risk population. Population size estimates document the existence and magnitude of the populations at risk of HIV and assist in the efficient allocation of prevention measures. In low-level and concentrated HIV epidemics, establishing the size of various at-risk populations has been identified as one of the greatest difficulties in developing estimates of HIV prevalence (1). Although there has been extensive work in the wildlife biology and fisheries fields to estimate the size of animal populations, there are fewer papers in the epidemiology literature using such methods (2–6). The goal of this study was to develop an unbiased estimator of the size of at-risk human populations that can be easily implemented in resource-limited settings.

Many methods have been used to estimate the size of at-risk human populations, but each method has limitations. For behaviors that are widespread in the general population, population survey methods provide robust estimates (7, 8). However, most high-risk behaviors for HIV are not prevalent in the general population and are stigmatized, which often results in underreporting. Furthermore, at-risk populations are often hidden and are not likely to be reached by such surveys. Ethnographic surveys that use nominative techniques, snowball sampling, or privileged access interviews use a small, accessible fraction of a larger concealed population to identify others exhibiting similar risk behavior. Although these methods are convenient in accessing hard-to-reach populations, generalizability of their findings is limited (9). Multiplier methods such as those associated with respondent-driven sampling use information from two overlapping sources, for example, an institution in contact with

Correspondence to Dr. Jacqueline E. Tate, Epidemiology Branch, Division of Viral Diseases, National Center for Immunizations and Respiratory Diseases, Centers for Disease Control and Prevention, 1600 Clifton Road NE, MS A47, Atlanta, GA 30333 (e-mail: jqt8@cdc.gov).

the target population and the population itself (8). This technique requires good institutional record keeping, data from institutions and populations that overlap, and clear definitions of populations, time reference periods, and catchment areas. Capture-recapture methods have been successful in estimating the size of hard-to-reach populations, but their usefulness is limited by the availability of independent, uncorrelated samples in which individuals in the target population have an equal probability of being captured, by the ability to identify individuals as captures and recaptures, and by difficulties in standardizing the definition of the target population (2, 10).

Laska, Meisner, and Siegel (LMS) (11) derived an unbiased estimator of population size based on a single sample. This method assumes that individuals in the population of interest appear on K lists, only one of which is observable. The population of interest is the observable number of distinct individuals on the Kth list plus the unobservable number of distinct individuals on the preceding $K-1$ lists who did not appear on list K. Individuals appear on the lists by engaging in a well-defined activity one or more times during a specified time period, such as K days. To implement this procedure, a survey of either all or a sample of all individuals engaging in the activity on the Kth day is conducted. The selected individuals are asked when they last engaged in the activity of interest. Based on this information, an unbiased estimator of the size of the target population during the K-day period can be obtained. LMS used this procedure to estimate the number of distinct individuals who received services from a mental health provider during a 52-week period (11, 12).

This article extends the LMS estimator from estimating the size of a population at a single venue to estimating the size of a population dispersed throughout multiple venues or units using two- and three-stage sampling. An application of the extended estimators using two- and three-stage sampling designs is presented to estimate the number of individuals who socialize over a 4-week period at public venues where transmission of HIV and other sexually transmitted infections is likely to occur. Mathematical details are given in the two appendices.

METHODS

Estimating population size in a single unit

Let t denote the size of a population of individuals engaging in the activity of interest one or more times during a K-day period (the time period of interest) ending in day K. Using maximum-likelihood estimation, LMS derive an unbiased estimator of t given by

$$\hat{t} = \frac{1}{\gamma}\sum_{h=1}^{K} h x_h, \qquad (1)$$

where γ is the known sampling fraction on the Kth day; x_h denotes the number of individuals in the sample on the survey day (day K) who last engaged in the activity h days before K for $h = 1, \ldots, K-1$; and x_K denotes the number of individuals in the sample who did not engage in the activity during the $K-1$ days before K. If all individuals, rather than a sample, are included, then the sampling fraction, γ, is equal to 1 and $\sum_{h=1}^{K} x_h$ is the total population of the unit on day K.

A sufficient condition for the LMS estimator $\hat{t}$ to be unbiased for t is given by the following (condition 1):

> *The probability of engaging in the activity of interest on the* K*th day and the (*K−h*)th day and on no days in between equals the average probability of this event taken over the* (K − *1* − h) *previous such events for* h = *0, . . . ,* K − *2.*

This condition ensures that data are collected during a typical period of time when the activity occurs. Further discussion of sufficient conditions for the LMS estimator to be unbiased is provided by Laska et al. (11).

The variance of the LMS estimator given in equation 1 is

$$\mathrm{var}(\hat{t}) = \frac{1}{\gamma^2}\sum_{h=1}^{K}(h^2-h)E(x_h) = \frac{1}{\gamma^2}\sum_{h=1}^{K}h^2E(x_h) - t. \qquad (2)$$

If condition 1 is assumed, then an unbiased estimator of $\mathrm{var}(\hat{t})$ is

$$\hat{\mathrm{var}}(\hat{t}) = \frac{1}{\gamma^2}\sum_{h=1}^{K}(h^2-h)x_h. \qquad (3)$$

Estimating population size in multiple units by using two-stage sampling

The LMS estimator of population size can be extended to two-stage sampling designs to estimate the size of a population dispersed throughout multiple units. In a two-stage sampling without replacement design, a sample of primary units is selected and then a sample of secondary units is chosen from each of the selected primary units (13). For example, the total number of individuals who socialize at public venues within a city can be determined by selecting a sample of venues within the city and then interviewing a sample of individuals socializing at the selected venues on day K about their frequency of visiting the venue.

Let N be the number of primary units in the population. For $i = 1, \ldots, N$ let t_i be the size of a population of secondary units in the ith primary unit that engage in the activity of interest one or more times during a K-day period ending on day K. Assuming that each secondary unit engages in the activity of interest at only one primary unit during the K-day period, then $\sum_{i=1}^{N} t_i = t$. Let n be the number of primary units sampled without replacement, M_i the number of secondary units in the ith sampled primary unit on day K, and m_i the number of secondary units selected without replacement from the ith sampled primary unit on day K, for $i = 1, \ldots, n$. An unbiased estimator of the total population during a K-day period at the ith primary unit in the sample is

$$\hat{t}_i = \frac{1}{\gamma_i}\sum_{h=1}^{K} h x_{ih} = \frac{M_i}{m_i}\sum_{h=1}^{K} h x_{ih}, \qquad (4)$$

where $\gamma_i = m_i/M_i$ is the known sampling fraction in the ith primary unit for $i = 1, \ldots, n$; x_{ih} is the number of secondary units in the sample from the ith primary unit who last engaged in the behavior of interest h days before K for $h = 1, \ldots, K{-}1$; and x_{iK} denotes the number of secondary units in the sample from the ith primary unit who did not engage in the behavior of interest during the $K{-}1$ days before K. An unbiased estimator of the population total during a K-day period is

$$\hat{t} = \frac{N}{n}\sum_{i=1}^{n}\hat{t}_i = \frac{N}{n}\sum_{i=1}^{n}\left(\frac{M_i}{m_i}\sum_{h=1}^{K} h x_{ih}\right). \tag{5}$$

A proof that this estimator is unbiased assuming that condition 1 holds for each primary unit is provided in appendix 1.

The variance of the estimator of the population total during a K-day period is

$$\mathrm{var}(\hat{t}) = N(N-n)\frac{\sigma_u^2}{n} + \frac{N}{n}\sum_{i=1}^{N}\sigma_i^2, \tag{6}$$

where

$$\sigma_u^2 = \frac{\sum_{i=1}^{N}\left(t_i - \frac{t}{N}\right)^2}{N-1} \tag{7}$$

and, for $i = 1, \ldots, N$,

$$\sigma_i^2 = \frac{1}{\gamma_i^2}\sum_{h=1}^{K}(h^2 - h)E(x_{ih}\,|\,s_1), \tag{8}$$

where s_1 is the sample of primary units. The first term on the right of the equality in equation 6 is the variance that would be obtained if every secondary unit in a selected primary unit were observed, that is, if the t_is were known for $i = 1, \ldots, n$. The second term contains variance due to estimating t_is from a subsample of secondary units within the selected primary units. An unbiased estimator of equation 6 is

$$\widehat{\mathrm{var}}(\hat{t}) = N(N-n)\frac{s_u^2}{n} + \frac{N}{n}\sum_{i=1}^{n}s_i^2, \tag{9}$$

where

$$s_u^2 = \frac{\sum_{i=1}^{n}\left(\hat{t}_i - \frac{\hat{t}}{N}\right)^2}{n-1} \tag{10}$$

and, for $i = 1, \ldots, n$,

$$s_i^2 = \frac{1}{\gamma_i^2}\sum_{h=1}^{K}(h^2 - h)x_{ih}.$$

Refer to appendix 2 for derivation of this two-stage sampling variance and its unbiased estimator. Based on maximum-likelihood results from LMS, approximate large-sample confidence intervals for t can be calculated by

$$\hat{t} \pm z_{1-\alpha/2}\widehat{\mathrm{var}}(\hat{t}),$$

where $z_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of the standard normal distribution.

If all primary units were selected, that is, if $N = n$, then the estimator for the total population size is simply the sum of the individual primary unit totals. The corresponding variance is the sum of the variances for each of the primary units because the primary units are independent. In particular, the first term to the right of the equality in equation 6 becomes 0 and the variance of $\hat{t}$ becomes

$$\mathrm{var}(\hat{t}) = \sum_{i=1}^{N}\sigma_i^2 = \sum_{i=1}^{N}\left\{\frac{1}{\gamma_i^2}\sum_{h=1}^{K}(h^2 - h)E(x_{ih}\,|\,s_1)\right\}$$

and an unbiased estimator of the variance $\hat{t}$ is

$$\widehat{\mathrm{var}}(\hat{t}) = \sum_{i=1}^{N}s_i^2 = \sum_{i=1}^{N}\left\{\frac{1}{\gamma_i^2}\sum_{h=1}^{K}(h^2 - h)x_{ih}\right\}.$$

Estimating population size in multiple units by using three-stage sampling

To estimate the population size at multiple venues using three-stage sampling, the unbiased estimator of population size from LMS can be further extended by following the pattern presented above for two-stage sampling. In a three-stage sampling without replacement design, a sample of primary units is selected, then a sample of secondary units is chosen from each of the selected primary units, and finally a sample of tertiary units is chosen from each selected secondary unit on day K. For example, in a large city where it is not possible to include the entire city in the survey, the city can be subdivided into administrative units. The total number of individuals who socialize at public venues within the city can be determined by first selecting a sample of administrative units, then choosing a sample of venues within the selected administrative units, and finally interviewing a sample of individuals socializing at the selected venues on day K to determine the frequency with which these individuals visit the venue.

Let L_{ij} be the number of tertiary units in the jth secondary unit of the ith primary unit on day K for $j = 1, \ldots, M_i$ and $i = 1, \ldots, N$. Let t_{ij} denote the size of the population of tertiary units in the jth secondary unit of the ith primary unit that engage in the activity of interest one or more times during a K-day period ending in day K. Assuming that each tertiary unit engages in the activity of interest at only one secondary unit during the K-day period, then $\sum_{i=1}^{N}\sum_{j=1}^{M_i} t_{ij} = t$. Again, let n be the number of primary units sampled without replacement, let m_i be the number of secondary units selected without replacement from the ith sampled primary unit, and let l_{ij} be the number of tertiary units selected from the jth secondary unit in the ith primary unit, for $j = 1, \ldots, m_i$ and $i = 1, \ldots, n$. An unbiased estimator of the total population during a K-day period at the jth secondary unit in the ith primary unit in the sample is

$$\hat{t}_{ij} = \frac{1}{\gamma_{ij}}\sum_{h=1}^{K}h x_{ijh} = \frac{L_{ij}}{l_{ij}}\sum_{h=1}^{K}h x_{ijh},$$

where $\gamma_{ij} = l_{ij}/L_{ij}$ is the known sampling fraction for tertiary units in the jth secondary unit of the ith primary unit on day K; x_{ijh} denotes the number of individuals (tertiary units) in the sample from the jth secondary unit of the ith primary unit who last engaged in the behavior of interest h days before K for $h = 1, \ldots, K{-}1$; x_{ijK} denotes the number of individuals in the sample from the jth secondary unit of the ith primary unit who did not engage in the behavior of interest during the $K{-}1$ days before K. An unbiased estimator of the population total in the ith primary unit in the sample during a K-day period is

$$\hat{t}_i = \frac{M_i}{m_i}\sum_{j=1}^{m_i}\hat{t}_{ij}.$$

Finally, an unbiased estimator of the population total during the K-day period is

$$\hat{t} = \frac{N}{n}\sum_{i=1}^{n}\hat{t}_i = \frac{N}{n}\sum_{i=1}^{n}\left\{\frac{M_i}{m_i}\sum_{j=1}^{m_i}\left(\frac{L_{ij}}{l_{ij}}\sum_{h=1}^{K}hx_{ijh}\right)\right\}. \tag{11}$$

The variance of the estimator of the population total during a K-day period is

$$\begin{aligned}\mathrm{var}(\hat{t}) = {} & N(N-n)\frac{\sigma_u^2}{n} + \frac{N}{n}\sum_{i=1}^{N}M_i(M_i-m_i)\frac{\sigma_i^2}{m_i} \\ & + \frac{N}{n}\sum_{i=1}^{N}\frac{M_i}{m_i}\sum_{j=1}^{M_i}\sigma_{ij}^2,\end{aligned} \tag{12}$$

where σ_u^2 is given in equation 7 and, for $i = 1, \ldots, N$,

$$\sigma_i^2 = \frac{\sum_{j=1}^{M_i}\left(t_{ij} - \frac{t_i}{M_i}\right)^2}{M_i - 1} \tag{13}$$

and, for $i = 1, \ldots, N$ and $j = 1, \ldots, M_i$,

$$\sigma_{ij}^2 = \frac{L_{ij}^2}{l_{ij}^2}\sum_{h=1}^{K}(h^2-h)E(x_{ijh}\,|\,s_1, s_2), \tag{14}$$

where s_1 is the sample of primary units and s_2 the sample of secondary units. The first term to the right of the equality in equation 12 is variance that would be obtained if every tertiary unit in a selected secondary unit and every secondary unit in a selected primary unit were observed, that is, if t_i were known for $i = 1, \ldots, n$. The second term contains variance that would be obtained if every tertiary unit in a selected secondary unit were observed, that is, if t_{ij} were known for $i = 1, \ldots, n$ and $j = 1, \ldots, m_i$. The third term contains variance due to estimating the t_{ij}s from a subsample of tertiary units within the selected secondary units. An unbiased estimator of the variance of $\hat{t}$ is obtained by replacing the population variances with the sample variances:

$$\begin{aligned}\widehat{\mathrm{var}}(\hat{t}) = {} & N(N-n)\frac{s_u^2}{n} + \frac{N}{n}\sum_{i=1}^{n}M_i(M_i-m_i)\frac{s_i^2}{m_i} \\ & + \frac{N}{n}\sum_{i=1}^{n}\frac{M_i}{m_i}\sum_{j=1}^{m_i}s_{ij}^2,\end{aligned} \tag{15}$$

where s_u^2 has the same form as equation 10 and, for $i = 1, \ldots, n$,

$$s_i^2 = \frac{\sum_{j=1}^{m_i}\left(\hat{t}_{ij} - \frac{\hat{t}_i}{M_i}\right)^2}{m_i - 1}, \tag{16}$$

and, for $i = 1, \ldots, n$ and $j = 1, \ldots, m_i$,

$$s_{ij}^2 = \frac{L_{ij}^2}{l_{ij}^2}\sum_{h=1}^{K}(h^2-h)x_{ijh}.$$

APPLICATION

Public venues where individuals engage in risky sexual and injection drug use behaviors that facilitate the transmission of HIV and other sexually transmitted infections provide a feasible and stable location to introduce prevention programs. These public venues are often the last place to reach individuals before they engage in risky behaviors. To design a venue-based prevention program, it is important to know the number of people socializing at the targeted venues. The Priorities for Local AIDS Control Efforts (PLACE) protocol provides a rapid, systematic method to identify areas likely to have a high incidence of risky sexual and drug-use behaviors. Within these areas, the PLACE method identifies specific public venues where HIV/acquired immunodeficiency syndrome prevention programs can be located to reach these populations (14). The PLACE method collects information through a series of cross-sectional surveys. First, community informant interviews are conducted to create a list of public venues within the city where people meet new sexual partners and/or injection drug users socialize. Next, these venues are visited to verify their existence and to obtain characteristics of the venues and their patrons important for developing a prevention program. Finally, a sample of venues is selected, and, at each of the chosen venues, a sample of individuals socializing is interviewed. PLACE studies were performed in Almaty, Kazakhstan (population 1.5 million) and Osh, Kyrgyzstan (population 200,000) in 2003.

To estimate the number of individuals who visited targeted venues in Osh during a 4-week (28-day) period, two-stage sampling was used. In this example, primary units are public venues within the city where risky sexual and drug-use behaviors occur; secondary units are individuals socializing at these venues. The activity of interest is visiting the venue. In an interview, the sampled respondents were asked how many days out of the past 7 they had visited the venue as well as the frequency with which they had visited the venue during the past 4 weeks. This information was combined to estimate the number of days prior to the interview day that each individual in the sample last visited the venue. If an individual reported visiting the venue multiple times during a week, the number of days since the last visit was calculated as seven divided by the number of days that the individual visited the venue in the past week. If the individual reported visiting the venue less than once a week, then the number of days since the last visit was estimated to be 14 days for individuals who reported visiting the venue “2–3 times a month” and 28 days for people who visited no more than once a month.

A total of $N = 237$ unique venues were identified in Osh as places where people meet new sexual partners and/or injection drug users socialize, and $n = 74$ venues were randomly selected in the first stage of sampling. The number of interviews performed at each venue depended on the total population size of the venue during a typical busy time. Ten individuals were interviewed at small venues (defined as having <20 men socializing at busy times), 20 individuals at medium-sized venues (20–49 men socializing at busy times), and 30 individuals at large venues (≥50 men socializing at busy times). Interviewers were instructed to choose potential respondents in a manner that minimized selection bias. An estimate of the total population size for each venue at the time of sampling was obtained from an interview with a representative at each venue. This individual was asked to estimate the total number of men and women socializing at the venue during a typical busy time. The conditions for unbiasedness in the LMS estimator were likely met because all interviews were conducted at "typical" busy times at the venues.

Using the unbiased estimator for two-stage sampling given by equation 5, the estimated number of individuals socializing at targeted venues during busy times over a 4-week period in Osh is 56,171, with an estimated variance of approximately 4.09×10^7obtained from equation 9. An asymptotic 95 percent confidence interval for the estimate is 43,634, 68,708. The variance due to sampling of secondary units within primary units is 3.1×10^6, and the variance due to sampling of primary units is $3.78 \times 10^{\ 7}$, indicating that the sampling of secondary units resulted in only a 4 percent increase in the estimated standard error compared with single-stage sampling.

To estimate the population size at targeted venues in Almaty, the large size of the city made it unfeasible to include the entire city in the study. Thus, three-stage sampling was needed to estimate the size of the socializing population at targeted venues during a 4-week period. The city was divided into $N = 72$ administrative units, $n = 15$ of which were randomly selected for the study. In this example, the primary units are the administrative units, the secondary units are the public venues, and the tertiary units are individuals socializing at the venues. Community informants in the $n = 15$ administrative units randomly selected in the first sampling stage identified 252 venues in these units, $\sum M_i = 149$ of which were identified as feasible locations for prevention programs. Of these $\sum M_i = 149$ venues, $\sum m_i = 37$ venues were randomly selected in the second stage of sampling. Each of the selected venues was visited by interviewers, and a sample of individuals was interviewed at each venue by using the same sampling procedure as in Osh. Using the estimator for three-stage sampling given by equation 11, the estimated number of individuals socializing at targeted venues during busy times over a 4-week period in Almaty is 262,557, with an estimated variance of approximately 8.60×10^9 obtained from equation 15. An asymptotic 95 percent confidence interval is 80,815, 444,319. The variance contribution due to primary unit sampling is 8.17×10^9, the variability due to secondary unit sampling is 3.13×10^8, and that due to tertiary unit sampling is 1.21×10^8.

DISCUSSION

Extension of the LMS estimator to incorporate survey designs with two- and three-stage sampling schemes increases the situations in which the estimator can be used. Populations of interest are not always contained within a single venue, nor is it often operationally feasible to visit all venues and interview all individuals at these venues. Although the two- and three-stage sampling designs slightly increase variability in the population size estimate, these designs enable surveys to be completed more quickly and reduce survey costs. For example, in Almaty, sampling of secondary and tertiary units combined contributed only 5 percent of the total variability. Because individuals within these secondary and tertiary units appear to be more similar within than across primary units, increasing the number of primary units sampled and decreasing the sample size of secondary and tertiary units will enable more precise estimates while not significantly increasing time or cost.

Extension of the LMS estimator for use with two- and three-stage sampling requires an additional assumption over single-stage sampling: individuals must frequent only one venue during the study period. Visits to multiple venues during the study period will result in an overestimate of population size. In this situation, the estimator can be thought of as an upper bound of the population size. To minimize the effect of multiplicity, this estimator should be used to estimate population sizes over only short time frames. Alternatively, one can relax this assumption by collecting additional information from respondents about other venues they frequent during the time period of interest (15). In particular, suppose in a separate study that we draw a simple random sample from the population of t distinct individuals who visited at least one venue during the study period. For each individual in the sample, suppose we record the number of primary units (i.e., venues) visited during the study period and let $\bar{W}$ denote the sample average number of venues visited. Then, $\hat{t}/\bar{W}$ is an asymptotically unbiased estimator of t (15). For example, if a simple random sample of individuals in Osh yields $\bar{W} = 2$, then the estimated number of individuals socializing at the targeted venues during a typical 4-week period would be 28,086, or half the number estimated under the assumption that each individual visits only one venue. Adjusting for visits to multiple venues will further increase the situations in which this estimator can be used to determine the size of venue-based populations.

Other limitations in interpreting our population estimates stem from the type of data collected, since estimating the population size of individuals socializing at these venues was not the primary objective of the study. First, the total number of individuals at each venue at the time of sampling is unknown but was estimated based on the typical number at the venue during a busy time. Formally incorporating this uncertainty in the sampling fraction would increase the variance, although preliminary calculations (results not shown) indicate that the increase would be slight. Second, respondents were not asked directly when they last visited the venue. Instead, the day of their last visit was estimated from the frequency with which they reported visiting the venue. In future studies, proper study design and planning will enable the limitations encountered in this application to be avoided.

ACKNOWLEDGMENTS

This research was partially supported by National Institute of Allergy and Infectious Diseases grant 5 P30 AI50410-08 and the Population, Health, and Nutrition Office USAID through MEASURE Evaluation Project Cooperative Agreement HRN-A-00-97-00018-00.

Conflict of interest: none declared.

REFERENCES

1. Ramon JS, Alvarenga M, Walker N, et al. Estimating HIV/AIDS prevalence in countries with low-level and concentrated epidemics: the example of Honduras. AIDS 2002;16(suppl 3): S18–22.
2. Seber GAF. The estimation of animal abundance and related parameters. 2nd ed. Caldwell, NJ: Blackburn Press, 2002.
3. Seber GAF. A review of estimating animal abundance II. Intl Stat Rev 1992;60:129–66.
4. Schwarz CJ, Seber GAF. Estimating animal abundance: review III. Stat Sci 1999;14:427–56.
5. Capture-recapture and multiple-systems estimation: I. History and theoretical development. International Working Group for Disease Monitoring and Forecasting. Am J Epidemiol 1995; 142:1047–58.
6. Capture-recapture and multiple-record systems estimation: II. Applications in human diseases. International Working Group for Disease Monitoring and Forecasting. Am J Epidemiol 1995;142:1059–68.
7. Pisani E. Estimating the size of populations at risk for HIV: issues and methods. Arlington, VA: USAID/IMPACT/Family Health International, 2002.
8. Archibald CP, Jayaraman GC, Major C, et al. Estimating the size of hard-to-reach populations: a novel method using HIV testing data compared to other methods. AIDS 2001; 15(suppl 3):S41–8.
9. Griffiths P, Gossop M, Powis B, et al. Reaching hidden populations of drug users by privileged access interviewers: methodological and practical issues. Addiction 1993;88: 1617–26.
10. Hook EB, Regal RR. Capture-recapture methods in epidemiology: methods and limitations. Epidemiol Rev 1995;17: 243–64.
11. Laska EM, Meisner M, Siegel C. Estimating the size of a population from a single sample. Biometrics 1988;44: 461–72.
12. Laska E, Lin S, Meisner M. Estimating the size of a population from a single sample: methodology and practical issues. J Clin Epidemiol 1997;50:1143–54.
13. Thompson SK. Sampling. New York, NY: John Wiley & Sons, 1992.
14. Weir SS, Pailman C, Mahalela X, et al. From people to places: focusing AIDS prevention efforts where it matters most. AIDS 2003;17:895–903.
15. Laska EM, Meisner M, Wanderling JA, et al. Estimating population size when duplicates are present. Stat Med 1996; 15:1635–46.

APPENDIX 1

Proof that the Two- and Three-Stage Sampling Estimators Are Unbiased

From LMS, we know that the expectation of $\hat{t}_i$ given by equation 4 conditional on a sample s_1 of primary units equals t_i assuming that condition 1 holds for each primary unit; that is, $E(\hat{t}_i|s_1) = t_i$. To obtain the expected value of $\hat{t}$ given by equation 5 over all possible samples of primary units, let

$$z_i = \begin{cases} 1 & \text{if the } i\text{th primary unit is in the sample} \\ 0 & \text{otherwise} \end{cases}$$

such that $E(z_i) = \frac{n}{N}$, the inclusion probability for a primary unit under simple random sampling.

Then, the expectation of $\hat{t}$ given in equation 5 is

$$E(\hat{t}) = E\{E(\hat{t}\,|s_1)\} = E\left\{E\left(\frac{N}{n}\sum_{i=1}^{n}\hat{t}_i\,|\,s_1\right)\right\}$$

$$= E\left\{\frac{N}{n}\sum_{i=1}^{n}t_i\right\} = E\left(\frac{N}{n}\sum_{i=1}^{N}z_i t_i\right) = t.$$

The proof of unbiasedness for the three-stage sampling estimator follows a similar pattern wherein the conditional expectation of the estimator is further broken down as

$$E(\hat{t}) = E\left[E\{E(\hat{t}\,|\,s_1, s_2)\,|\,s_1\}\right],$$

where s_2 denotes the sample of secondary units.

APPENDIX 2

Derivation of the Variance and Its Unbiased Estimator using a Two-Stage Sampling Estimator

To derive the variance of $\hat{t}$ given in equation 5, we use

$$\text{var}(\hat{t}) = \text{var}\{E(\hat{t}\,|s_1)\} + E\{\text{var}(\hat{t}|s_1)\}. \qquad \text{(A1)}$$

Because of the simple random sampling of primary units without replacement at the first stage, the first term to the right of the equality in equation A1 equals

$$\text{var}\{E(\hat{t}\,|s_1)\} = \text{var}\left(\frac{N}{n}\sum_{i=1}^{n}t_i\right) = N(N-n)\frac{\sigma_u^2}{n}, \qquad \text{(A2)}$$

where σ_u^2 is given by equation 7. For the second term to the right of the equality in equation A1,

$$\text{var}(\hat{t}\,|\,s_1) = \text{var}\left(\frac{N}{n}\sum_{i=1}^{n}\hat{t}_i\,|\,s_1\right) = \left(\frac{N}{n}\right)^2\sum_{i=1}^{n}\text{var}(\hat{t}_i\,|\,s_1)$$

$$= \frac{N^2}{n^2}\sum_{i=1}^{n}\sigma_i^2,$$

where σ_i^2 is given in equation 8, and the last equality follows from LMS. Therefore,

$$E\{\mathrm{var}(\hat{t}\,|\,s_1)\} = E\left\{\frac{N^2}{n^2}\sum_{i=1}^{N} z_i\sigma_i^2\right\} = \frac{N}{n}\sum_{i=1}^{N}\sigma_i^2. \tag{A3}$$

Combining equations A2 and A3 according to equation A1 yields equation 6.

To see that equation 9 is an unbiased estimator of equation 6, first note that s_u^2 given in equation 10 can be rewritten as

$$s_u^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n}\hat{t}_i^2 - \frac{\hat{t}^2 n}{N^2}\right). \tag{A4}$$

Next, note

$$\begin{aligned} E\left(\sum_{i=1}^{n}\hat{t}_i^2\right) &= E\left\{E\left(\sum_{i=1}^{n}\hat{t}_i^2\,|\,s_1\right)\right\} \\ &= E\left(\sum_{i=1}^{n}\left[\mathrm{var}(\hat{t}_i\,|\,s_1) + \{E(\hat{t}_i\,|\,s_1)\}^2\right]\right) \\ &= E\left(\sum_{i=1}^{n}\sigma_i^2 + \sum_{i=1}^{n}t_i^2\right), \end{aligned}$$

implying

$$\begin{aligned} E\left(\sum_{i=1}^{n}\hat{t}_i^2\right) &= E\left(\sum_{i=1}^{N}z_i\sigma_i^2 + \sum_{i=1}^{N}z_i t_i^2\right) \\ &= \frac{n}{N}\left(\sum_{i=1}^{N}\sigma_i^2 + \sum_{i=1}^{N}t_i^2\right). \end{aligned} \tag{A5}$$

In addition,

$$\begin{aligned} E(\hat{t}^2) &= \mathrm{var}(\hat{t}) + \{E(\hat{t})\}^2 \\ &= \frac{N(N-n)}{n}\sigma_u^2 + \frac{N}{n}\sum_{i=1}^{N}\sigma_i^2 + t^2. \end{aligned} \tag{A6}$$

Together, equations A4, A5, and A6 imply

$$\begin{aligned} E(s_u^2) &= \frac{n}{(n-1)N}\left(\sum_{i=1}^{N}\sigma_i^2 + \sum_{i=1}^{N}t_i^2\right) \\ &\quad - \frac{n}{(n-1)N^2}\left(\frac{N(N-n)}{n}\sigma_u^2 + \frac{N}{n}\sum_{i=1}^{N}\sigma_i^2 + t^2\right) \\ &= \frac{1}{N}\sum_{i=1}^{N}\sigma_u^2 + \frac{(N-1)n}{(n-1)N}\left\{\frac{1}{N-1}\left(\sum_{i=1}^{N}t_i^2 - \frac{t^2}{N}\right)\right\} \\ &\quad + \frac{(N-n)}{N(n-1)}\sigma_u^2 = \frac{1}{N}\sum_{i=1}^{N}\sigma_i^2 + \sigma_u^2, \end{aligned} \tag{A7}$$

where the last equality uses the fact that $\sigma_u^2 = \frac{1}{N-1}\left(\sum_{i=1}^{N} t_i^2 - \frac{t^2}{N}\right)$. Next, note that

$$\begin{aligned} E\left(\sum_{i=1}^{n}s_i^2\right) &= E\left\{E\left(\sum_{i=1}^{n}s_i^2|s_1\right)\right\} \\ &= E\left\{\sum_{i=1}^{n}\frac{M_i^2}{m_i^2}\sum_{h=1}^{K}(h^2-h)E(x_{ih}|s_1)\right\} \\ &= E\left\{\sum_{i=1}^{N}z_i\sigma_i^2\right\} = \frac{n}{N}\sum_{i=1}^{N}\sigma_i^2. \end{aligned} \tag{A8}$$

Together, equations A7 and A8 imply that the expected value of equation 9 equals equation 6; that is,

$$\begin{aligned} E\{\hat{\mathrm{var}}(\hat{t})\} &= \frac{N(N-n)}{n}\sigma_u^2 + \frac{(N-n)}{n}\sum_{i=1}^{N}\sigma_i^2 \\ &\quad + \sum_{i=1}^{N}\sigma_i^2 = \mathrm{var}(\hat{t}). \end{aligned}$$

For three-stage sampling, the conditional variance in equation A1 can be further broken down as

$$\mathrm{var}(\hat{t}\,|\,s_1) = \mathrm{var}\{E(\hat{t}\,|\,s_1,s_2)\,|\,s_1\} + E\{\mathrm{var}(\hat{t}\,|\,s_1,s_2)\,|\,s_1\},$$

and the derivation of the variance and its unbiased estimator for three-stage sampling follows the same pattern as for two-stage sampling.