INTERNATIONAL ATOMIC ENERGY AGENCY
UNITED NATIONS EDUCATIONAL, SCIENTIFIC AND CULTURAL ORGANIZATION
INTERNATIONAL CENTRE FOR THEORETICAL PHYSICS
I.C.T.P., P.O. BOX 586, 34100 TRIESTE, ITALY, CABLE: CENTRATOM TRIESTE

UNITED NATIONS INDUSTRIAL DEVELOPMENT ORGANIZATION

INTERNATIONAL CENTRE FOR SCIENCE AND HIGH TECHNOLOGY
c/o INTERNATIONAL CENTRE FOR THEORETICAL PHYSICS 34100 TRIESTE (ITALY) VIA GRIGNANO, 9 (ADRIATICO PALACE) P.O. BOX 586 TELEPHONE 040-224572 TELEFAX 040-224575 TELEX 460449 APH I

SMR.648 - 35

SECOND AUTUMN WORKSHOP ON MATHEMATICAL ECOLOGY

(2 - 20 November 1992)

"Using Mark-Recapture Methodology to Estimate the Size of a Population at Risk for Sexually Transmitted Diseases"

C. Castillo-Chavez
Biometrics Unit
Cornell University
Ithaca, NY 14853-7801
U.S.A.

These are preliminary lecture notes, intended only for distribution to participants.

USING MARK-RECAPTURE METHODOLOGY TO ESTIMATE THE SIZE OF A POPULATION AT RISK FOR SEXUALLY TRANSMITTED DISEASES

GAIL RUBIN, DAVID UMBACH, SHWU-FANG SHYU AND
CARLOS CASTILLO-CHAVEZ
Biometrics Unit, Cornell University, 337 Warren Hall, Ithaca, New York 14850,
U.S.A.
607-255-5488

BU-1112-MB

March 1992

USING MARK-RECAPTURE METHODOLOGY TO ESTIMATE THE SIZE OF A POPULATION AT RISK FOR SEXUALLY TRANSMITTED DISEASES

SUMMARY

To study the spread of sexually transmitted diseases (STDs) using social / sexual mixing models, one must have quantitative information about sexual mixing. An unavoidable complication in gathering such information by survey is that members of the surveyed population will almost certainly have sexual contacts outside that population. The number of these outsiders may be substantial and, hence, important for the modelling process. In this paper, we develop a mark-recapture model for estimating the size of the population at risk for contracting a STD due to direct sexual contact with a specified population targeted by a survey. This mark-recapture methodology provides a reliable method of estimating the number of outsiders. Because not everyone in the targeted population may be sexually active, the size of the sexually active subset, used as the number marked in our tag-recapture formulation, must be estimated, which introduces extra variability. We derive an estimator of the variance of the estimated total number at risk that accounts for this extra variability and an expression for the bias of that estimator. We extend the methodology to stratified surveys and illustrate its use with data collected from a population of university undergraduates to estimate sexual mixing parameters of a deterministic model of the spread of STDs.

Using mark-recapture methodology to estimate the size of a population at risk for sexually transmitted diseases

BU-1112-MB[1]

March 1992

Gail Rubin, David Umbach, Shwu-Fang Shyu and Carlos Castillo-Chavez

ABSTRACT

To study the spread of sexually transmitted diseases (STDs) using social / sexual mixing models, one must have quantitative information about sexual mixing. An unavoidable complication in gathering such information by survey is that members of the surveyed population will almost certainly have sexual contacts outside that population. The number of these outsiders may be substantial and, hence, important for the modelling process. In this paper, we develop a mark-recapture model for estimating the size of the population at risk for contracting a STD due to direct sexual contact with a specified population targeted by a survey. This mark-recapture methodology provides a reliable method of estimating the number of outsiders. Because not everyone in the targeted population may be sexually active, the size of the sexually active subset, used as the number marked in our tag-recapture formulation, must be estimated, which introduces extra variability. We derive an estimator of the variance of the estimated total number at risk that accounts for this extra variability and an expression for the bias of that estimator. We extend the methodology to stratified surveys and illustrate its use with data collected from a population of university undergraduates to estimate sexual mixing parameters of a deterministic model of the spread of STDs.

[1] In the Biometrics Unit Technical Report Series, Cornell University, Ithaca, NY 14853.

1. INTRODUCTION

Public health workers have long recognized the importance of social and sexual interactions in the spread of sexually transmitted diseases (STDs), especially HIV/AIDS. Quantifying that understanding in ways that allow modelling of how different rates of mixing between subgroups in a population alter the time course of an epidemic, however, has a more recent origin. New statistical and mathematical models of the spread of HIV through a population have incorporated social and sexual factors[1] or interactions.[2-5] In general terms, one can examine sexual mixing by dividing a population into relevant subgroups and examining rates of sexual contact within and between subgroups. Mathematical models of disease spread allow the investigator to vary the mixing parameters and examine the projected course of the epidemic. Reliable estimates of the values of mixing parameters and of the sizes of the subgroups, which one must obtain by gathering data on the sexual behavior of individuals, are crucial to realistic modelling of the process of pair formation and, hence, the spread of STDs.

Our particular interest is in estimating the size of a population at risk for contracting a STD due to direct sexual contact with a specified population. The specified population, whose size is known, is targeted by a survey to gather information about the number of sexual partners its individuals have; the reported sexual partners also are classified as members of the surveyed population or as outsiders. The outsiders form a subgroup of the sexually interacting population which we cannot survey directly. With surveys in which the group targeted gives information about sexual contact with outsiders, mark-recapture methodology provides a reliable means of estimating both the size of the outsider subgroup and its variance. In this paper, we formulate a mark-recapture model appropriate for such data and use it to estimate the size of the population at risk, and consequently the size of the subgroup of outsiders.

Because not everyone in the population targeted by the survey may be sexually active, the estimation problem is complicated by the need to estimate the size of the sexually active subgroup of this target population. The population at risk contains all sexually active individuals in the population targeted by the survey plus all their sexual partners who are outsiders. (We want to emphasize that our term "population at risk" refers only to persons in *direct* contact with the target population; it does not refer to individuals who may eventually contract a disease through a chain of contacts leading back to the target population.) We develop methodology, based on Bailey's mark-recapture model[6], to estimate the size of the population at risk. We illustrate our methodology with survey data, from a university undergraduate population, collected to estimate sexual mixing parameters for modelling the spread of sexually transmitted diseases such as HIV/AIDS.[7]

The application of mark-recapture methodology (synonymous with capture-recapture and tag-recapture methology) to epidemiology is not new, although it is uncommon. Using merged hospital lists of patients having a certain trait that is rare in the population at large, Wittes[8] applied capture-recapture methods to estimate the size of the population having that trait. Goldberg and Wittes[9] used similar methods to estimate the number of false negatives in medical screening for early detection of breast cancer. When we can apply mark-recapture methodology to public health or survey data to estimate the size of the population at risk for a sexually transmitted disease, the resulting estimators are design-based rather than model-based, in the sense that they do not rely on a probabilistic model for the population whose size we wish to estimate, but depend instead on the sampling design. Therefore, mark-recapture population estimates can provide an independent benchmark against which to compare estimates based on different probabilistic models.

In Section 2, we describe the survey conducted by Crawford, Schwager and Castillo-Chavez[7], which we use in Section 7 to illustrate the methodology developed in this paper. We briefly review Bailey's model[6] and estimators in Section 3 and then, in Section 4, we give a mark-recapture framework, involving two stages of estimation, for use with survey data. We also examine the assumptions needed for valid inference under this formulation. In Section 5, we derive estimators of the number of people of each sex at risk for disease, and the variance of that estimator. We consider the approximate bias of these estimators and also derive estimators of the number of outsiders and its variance. Finally, in Section 6, we extend the methodology to situations in which the primary strata (sexes) are stratified further.

2. CORNELL UNDERGRADUATE SOCIAL AND SEXUAL PATTERNS SURVEY

Crawford, Schwager and Castillo-Chavez[7] conducted a survey (CUSSP) at Cornell University, Ithaca, NY, USA in the fall of 1989 that solicited information about sexual and social behavior and drug and alcohol use during a specific two month period, from a stratified random sample of Cornell undergraduates (stratified by sex and class year). Unlike the survey conducted by Stigum et al[5], designed specifically for estimating the parameters required for modelling HIV/AIDS in the heterosexual population of Norway, the CUSSP survey had multiple objectives and the sampling frame was limited to the university. Of the 11,750 undergraduates registered, 1878 students were selected and 953 responded (response rate of 50.7 %). Only 502 students of the 953 who responded, however, received the version of the questionnaire (direct questionnaire) that solicited information required to estimate the parameters of social mixing models; the remaining 451 respondents completed the indirect version, which solicited information on dating. Crawford et al[7] evaluated the performance of the CUSSP survey in obtaining information on such sensitive subjects.

Each completed direct questionnaire contained the number of distinct sexual partners in the previous two month period in each of the following categories: Cornell undergraduates; staff, faculty or graduate students at Cornell; and people not affiliated with Cornell. The partners who were Cornell undergraduates were categorized further into class year (freshman, sophomore, junior and senior). A sexual partner was defined as one with whom the respondent engaged in penetrating vaginal or anal intercourse. To be considered sexually active, a student had to have at least one sexual partner in the two month period. We discuss later the effect of this definition of "sexually active" on the estimated number of people at risk. For purposes of modelling the spread of STDs in a heterosexual population, homosexuals and bisexuals were eliminated from the data base. Only one homosexual responded to the CUSSP survey, which is consistent with the lower bound estimates of Fay et al[10] that 3.3 - 6.2 % of males in a population are homosexual. Some of Fay's estimates[10] are unconditional, whereas others are conditional on being sexually active. For the CUSSP survey, the percentages homosexual of all males and of all sexually active males are 0.4 and 4.5 %, respectively.

Since the survey was conducted in late October and early November, the two month period in question covered time after the onset of the fall semester. (No resurvey of Cornell undergraduates was conducted.) Presumably, the information on sexual contacts should refer primarily to the Ithaca area, which houses Ithaca College as well as Cornell University, but potentially could extend to other nearby college communities (e.g., Cortland, Geneva, Aurora and Syracuse). During the academic year, college students comprise nearly half the population of Ithaca (city and town combined). Based on 1990 census figures and registrars' records from Cornell University and Ithaca College, the permanent resident population of Ithaca is approximately 25,000 , the Cornell student population is about 17,800 (12,000 undergraduates; 5,800 graduate students) and the Ithaca College student population is about 6,200 (100 of which are graduate students). Cornell undergraduates have ample opportunity for sexual contact with students of other colleges.

3. BAILEY'S MARK-RECAPTURE MODEL

Bailey[6] described a binomial model that he viewed as an approximation to the classic hypergeometric capture-recapture model for a single capture period after marking:

$$p\,(m_2 \mid n_1, n_2) \approx \binom{n_2}{m_2} (n_1 / N)^{m_2} \,\{1-(n_1 / N)\}^{n_2 - m_2}, \qquad (1)$$

where n_1 is the number of individuals marked at time 1, n_2 is the number of individuals captured at time 2, m_2 is the number of marked individuals captured at time 2, and N is the total population size. As Seber[11] (pp. 61, 565-566) points out, the model given above holds exactly for sampling with replacement, as when one merely observes rather than captures animals and different observers may see the same animal. As will become apparent shortly, sampling with replacement is the appropriate model for the CUSSP survey estimation problem.

Bailey[6] showed that the Lincoln-Petersen estimator for N, the maximum likelihood estimator under model (1), is biased. He suggested

$$\hat{N} = n_1\,(n_2+1)\,/\,(m_2+1), \qquad (2)$$

to estimate N and

$$\hat{v}\left(\hat{N}\right) = n_1^2\,(n_2+1)\,(n_2-m_2)\,/\,\{(m_2+1)^2\,(m_2+2)\} \qquad (3)$$

to estimate the variance of $\hat{N}$. These estimators have proportional biases of order $\exp\,(-\,n_1 n_2 / N)$ and $(n_1 n_2 / N)^2 \exp\,(-\,n_1 n_2 / N)$, respectively; the bias of $\hat{N}$ is negative whereas that of $\hat{v}\left(\hat{N}\right)$ is positive.[6]

4. MARK-RECAPTURE FRAMEWORK FOR CUSSP SURVEY

For the CUSSP survey, the population at risk consists of all sexually active Cornell undergraduates plus all their sexual partners who are outsiders. University undergraduates are highly mobile and may be highly sexually active, characteristics of a good vector for STDs. Accordingly, the network of sexual contacts may extend well beyond residents of the community in which they reside for the academic year. To model properly the spread of STDs throughout a population that includes college communities, we need a reliable estimate of the number of sexual partners who are outsiders. Our objective is to estimate the total number at risk and to use it to estimate the number of outsiders at risk. If properly applied, capture-recapture methodology provides good estimates of these parameters. In this section, we translate our problem into a mark-recapture scenario and examine the required assumptions.

The population targeted by the CUSSP survey is all Cornell undergraduates; Registrar's figures for Fall 1989 provide a frame. Responses to the direct questionnaire provide the number of people in the sample who meet our definition of "sexually active" and allow us to estimate the total number of sexually active undergraduates by using standard finite population methods. (This, of course, assumes that the response bias in the survey is negligible.) For the purpose of mark-recapture estimation, we take all sexually active Cornell undergraduates as the marked population.

The need to estimate the size of the marked population is a consequence of the stringent definition of a sexual partner used in the CUSSP survey. Although the survey gathered information on sexual activity other than intercourse, the data on sexual partners were limited to pairings that met the definition given in Section 2.

That is, the data on contacts were limited to two acts deemed the most likely, yet not the only ones, to result in heterosexual transmission of STDs or HIV. In addition, students who engaged in intercourse before or after the two month period do not contribute to the size of the population at risk, although they actually belong to it. Consequently, the definition of "sexually active" for the CUSSP survey may be too restrictive, and thus, lead to estimating only the number of people at greatest risk for sexually transmitted diseases. Furthermore, the estimate of the size of the population at risk depends on the time period covered by the survey (as do all tag-recapture population estimates). This dependence, however, is an unknown function of time, and we would require more information to model such dependence and incorporate it into estimation.[12]

The population at risk contains sexually active people, both Cornell undergraduates and outsiders, that is, both marked and unmarked individuals. The Cornell students surveyed identify themselves as marked or unmarked (i.e., sexually active or not), and the marked students classify their partners either as Cornell (marked) or as outsiders. By definition, the unmarked students surveyed contribute no partners. We have no information about outsiders who are not sexual partners of Cornell undergraduates. Since we access information about sexual partners *only* from the Cornell students surveyed, the students surveyed play the role of observers in mark-recapture studies in which "recapture" consists of sighting. For each student surveyed, the contacts reported are distinct sexual partners. Any two students surveyed, however, may share one or more sexual partners, either from the Cornell student pool or from the greater Ithaca area, so that the combined number may contain multiple counts of the same sexual partner. Thus, the surveyed students are sampling sexual partners *with replacement*. Consequently, the closed population, single mark release model, which is based on sampling with replacement[6], is an appropriate model for estimating the number of people at risk.

Having only a single recapture period (the two month period in question), we must use a closed population model. According to Pollack, Nichols, Brownie and Hines[13], the assumptions to consider are: [i-a] additions (births and immigrants) to the population are always unmarked; [i-b] deletions (deaths or emigrants) occur randomly with respect to marked and unmarked individuals; and [ii] marked and unmarked individuals have equal capture probabilities. Assumption [i-a] is likely true in our case, since new students are not admitted to Cornell during the two month period of interest. Assumption [i-b] also is likely true in our case, if the primary cause of deletion is illness rather than vacation or moving. The assumption that marked and unmarked individuals have equal capture probabilities is problematic. As Pollack et al[13] note (p. 10): "If capture probabilities are heterogeneous in each sample but independent from sample to sample, then no bias results." Unfortunately, we cannot assess the validity of equal accessibility to or acceptability of Cornell and non-Cornell partners, since our data come from a single recapture period. In Ithaca, however, partners from Ithaca College are equally accessible to Cornell students although not necessarily equally acceptable. One must recognize, however, that all capture-recapture studies suffer the flaw that one never knows whether the untrapped animals are as catchable as those trapped.[14] We need not worry about loss of marks or overlooking marks, a problem in many applications of mark-recapture to wildlife populations, because we do not expect that students hide their Cornell / non-Cornell affiliations.

5. BASIC ESTIMATION FOR HETEROSEXUAL POPULATIONS

In this section, we consider stratification of the population by a single factor, sex. In Section 6 we consider an extension to two way stratification of the marked population with each sex stratified further.

5.1. Estimating the number of sexually active undergraduates

Let the subscript i denote sex, having two levels (m = male, f = female). Let R_i denote the number of Cornell undergraduates of sex i registered in the Fall of 1989 (Registrar's figures) and T_i the number of those of sex i sexually active during the period in question. We denote the number of Cornell undergraduates in the sample and the corresponding number sexually active as r_i and t_i, respectively. Since the selection of students in the survey was by sampling without replacement from the Registrar's list, the number of sexually active students in the sample is a hypergeometric random variable:

$$p\,(t_i \mid r_i, T_i, R_i) = \binom{T_i}{t_i}\binom{R_i - T_i}{r_i - t_i} / \binom{R_i}{r_i},$$

which we can approximate by a binomial distribution

$$p\,(t_i \mid r_i, T_i, R_i) \approx \binom{r_i}{t_i}(T_i / R_i)^{t_i}\ (1 - (T_i / R_i))^{r_i - t_i}.$$

We can estimate T_i using the maximum likelihood estimator under the approximate binomial model as

$$\widehat{T}_i = R_i t_i / r_i \equiv R_i \widehat{\pi}_i ,$$

where $\widehat{\pi}_i$ estimates $\pi_i = T_i / R_i$, the probability of an individual of sex i in the surveyed population being sexually active. Since a random sample was independently drawn from each stratum, $\widehat{T}_i$ is an unbiased, consistent estimator of T_i under either the hypergeometric or the binomial model. Estimating the number marked by $\widehat{T}_i$ introduces extra variation into the estimation of the number at risk.

5.2. Estimating the size of the population at risk

Let N_i denote the total number of people of sex i at risk. Let y_i denote the total number of sexual contacts with sex i during the two month period reported by surveyed Cornell students; and let x_i represent the total number of sexual contacts with Cornell undergraduates of sex i during that period, with the difference $y_i - x_i$ being the number of sexual contacts of sex i that were outsiders (e.g., staff, graduate students or faculty from Cornell; people living in Ithaca but not affiliated with Cornell; friends not from the Ithaca area). For heterosexuals, the survey responses on number of partners given by men provide x_f and y_f whereas we use the number of women respondents to the survey that met the CUSSP definition of sexually active to estimate T_f -- the three values needed to estimate N_f. The sexes contribute the reverse data to estimate N_m. Again, we remind the reader, each student surveyed reports distinct sexual partners; however, any two students surveyed may share one or more sexual partners, either Cornell undergraduates or outsiders. Hence, the combined number of sexual contacts (y_i or x_i) may contain multiple counts of the same partner; and the combined count of partners represents a sample taken *with replacement* from the population at risk.

Using Bailey's model[6], equation (1) becomes

$$p\,(x_i \mid T_i, y_i) = \binom{y_i}{x_i}(T_i / N_i)^{x_i}\ \{1-(T_i / N_i)\}^{y_i - x_i}\,, \tag{4}$$

giving the exact probability of x_i contacts with the surveyed students conditional on T_i sexually active students and the total number of contacts, y_i. We must, however, substitute our estimate $\hat{T}_i$ for T_i.

Bailey's[6] estimator (2), re-expressed in our notation,

$$\hat{N}_i = T_i\,(y_i+1)\,/\,(x_i+1)\,, \tag{5}$$

is a nearly unbiased estimator of the number of sex i at risk, when the size of the marked group is known. The corresponding variance estimator, (3), is

$$\hat{v}\left(\hat{N}_i \mid y_i, T_i\right) = T_i^2\,(y_i+1)\,(y_i-x_i)\,/\,\{(x_i+1)^2\,(x_i+2)\}\,. \tag{6}$$

Adopting the point of view that all undergraduates are sexually active to some degree (i.e., $T_i = R_i$ for each i) removes the need to estimate the size of the sexually active subset. Then the size of the marked group is known, is equal to the Registrar's count, and we can use Bailey's model[6] and its estimators with R_i replacing T_i in (4) through (6). Under this assumption, the estimators given by (5) and (6) are conditional only on y_i, and the variance estimated by (6) is the appropriate estimate of the precision of $\hat{N}_i$ to report. This approach, however, does have the potential of overestimating the size of the population at risk. To proceed in a manner consistent with the CUSSP survey's definition of "sexually active," we estimate T_i.

The corresponding estimator of N_i, which uses the estimated size of the marked group,

$$\widetilde{N}_i = \widehat{T}_i\,(y_i+1)\,/\,(x_i+1)\,, \qquad (7)$$

is a nearly unbiased estimator also, with proportional bias of order

$$\left[1-\pi_i + \pi_i \exp\left\{-y_i\, R_i\,/\,(N_i\, r_i)\right\}\right]^{r_i} \equiv \left\{B\,(N_i)\right\}^{r_i}. \qquad (8)$$

The bias given by the LHS of (8), the moment generating function of a binomial random variable (t_i), arises from the expectation of the bias of $\widehat{N}_i$, when we use $\widehat{T}_i$ instead of T_i (see Section 3). Substituting $\widehat{T}_i$ in the estimator of the variance of $\widehat{N}_i$ given by (6), provides an estimator that is conditional on both y_i and $\widehat{T}_i$.

Seber[11] (p. 82) discusses the issue of a fixed vs. random sample size in a recapture period and points out that there is little difference between treating the sample size as fixed or random when the primary concern is estimation. Here, y_i corresponds to the random sample size, and the inference using $\widehat{v}\left(\widetilde{N}_i \mid y_i, \widehat{T}_i\right)$ is conditional on the total number of sexual contacts with members of sex i and on the estimated number of sexually active Cornell undergraduates of sex i. Because a given sexual partner can be reported by more than one of the students surveyed, the total number of contacts (y_i) can be even greater than N_i, thereby increasing the precision of the survey for N_i.[11]

We must include in the variance estimator the extra variation introduced into the estimation of N_i by using an estimate of the number of sexually active students ($\widehat{T}_i$). To acknowledge this extra variation in $\widetilde{N}_i$, we must use an estimator that is conditional only on y_i. Writing

$$\operatorname{var}\left(\widetilde{N}_i \mid y_i\right) = E\left\{\operatorname{var}\left(\widetilde{N}_i \mid y_i, \widehat{T}_i\right)\right\} + \operatorname{var}\left\{E\left(\widetilde{N}_i \mid y_i, \widehat{T}_i\right)\right\}, \qquad (9)$$

we find that the first term of the right-hand side (RHS) of (9) is approximately the ratio of

$$E\left[R_i^3\, y_i\, (y_i+1)^2\, t_i^3\, \left(N_i\, r_i^3\right)^{-1} \{1 - R_i\, t_i / (N_i\, r_i)\} \mid y_i\right] \tag{10}$$

to

$$E\left[\{y_i\, t_i\, R_i\, (N_i\, r_i)^{-1} + 1\}^4 \mid y_i\right]. \tag{11}$$

The expectations in (10) and (11) are quartic polynomials in π_i. Letting $a_i = R_i / (N_i\, r_i)$, expression (10) equals

$$\begin{aligned} A(\pi_i) = R_i^3\, y_i\, (y_i+1)^2 \left(N_i\, r_i^3\right)^{-1} r_i\, \pi_i \big[\{1 - a_i\} + \pi_i\, (r_i - 1)\{3 - 7\, a_i\} \\ + \pi_i^2\, (r_i - 1)(r_i - 2)\{1 - 6\, a_i\} - \pi_i^3\, (r_i - 1)(r_i - 2)(r_i - 3)\, a_i\big], \end{aligned} \tag{12}$$

and expression (11) equals

$$\begin{aligned} C(\pi_i) = \pi_i^4 r_i\, (r_i - 1)(r_i - 2)(r_i - 3)\, y_i^4 a_i^4 + 2\pi_i^3 r_i\, (r_i - 1)(r_i - 2)\, y_i^3 a_i^3\, \{3\, y_i a_i + 2\} \\ + \pi_i^2 r_i\, (r_i - 1)\;\; y_i^2 a_i^2 \left\{6\, (y_i a_i + 1)^2 + 1\right\} + \pi_i r_i y_i a_i\, (y_i a_i + 2)\{2 + y_i a_i\, (y_i a_i + 2)\} + 1\,. \end{aligned} \tag{13}$$

The expectation given in (10), a difference of scaled third and fourth moments of a binomial random variable, must be nonnegative for the approximate $\mathrm{var}\left(\tilde{N}_i \mid y_i\right)$ to be nonnegative. The expression given by (10), or equivalently (12), is nonnegative when

$$(N_i\, r_i / R_i)\left\{1 + 3\pi_i\, (r_i - 1) + \pi_i^2\, (r_i - 1)(r_i - 2)\right\} \geq 1 + \pi_i (r_i - 1)\left\{7 + 6\pi_i\, (r_i - 2) + \pi_i^2\, (r_i - 2)(r_i - 3)\right\}.$$

The second term of the RHS of (9) is a scaled difference of the moment generating functions of two binomial random variables, having the same probability of success but a different number of trials:

$$N_i^2\, \mathrm{var}\left[1 - \exp\{-\, y_i\, R_i\, t_i / (N_i\, r_i)\} \mid y_i\right] = N_i^2\left[\{B\, (N_i / 2)\}^{r_i} - \{B\, (N_i)\}^{2\, r_i}\right].$$

The RHS of the expression above is guaranteed to be nonnegative, since y_i, R_i, and r_i are all greater than zero. Its contribution, however, to the variance is negligible (< 10^{-6}) for the parameter configurations we examined. Substituting $\hat{\pi}_i$

and $\tilde{N}_i$ for π_i and N_i , respectively, into (12) and (13) yields

$$\hat{v}\left(\tilde{N}_i \mid y_i\right) = A(\hat{\pi}_i) / C(\hat{\pi}_i) + \tilde{N}_i^2 \left[\{B(\tilde{N}_i / 2)\}^{r_i} - \{B(\tilde{N}_i)\}^{2r_i}\right]. \tag{14}$$

Since $\tilde{N}_i$ is a ratio estimator, we can give only an approximation for the bias of $\hat{v}\left(\tilde{N}_i \mid y_i\right)$,

$$\operatorname{var}\left(\tilde{N}_i \mid y_i\right) - E\left[E\left\{\hat{v}\left(\tilde{N}_i \mid y_i, \hat{\pi}_i, x_i\right)\right\}\right]. \tag{15}$$

The iterated expectation in (15) is approximately equal to a ratio of 7th order polynomials in π_i (see Appendix I).

5.3. Estimating the number of outsiders of each sex

For the CUSSP survey, we wish to estimate the number of outsiders (N_i-T_i). Thus, an estimate of N_i-$T_i \equiv O_i$ is

$$\begin{aligned} \hat{O}_i &= \tilde{N}_i - \hat{T}_i = \{\hat{T}_i(y_i+1)/(x_i+1)\} - \hat{T}_i \\ &= \hat{T}_i[\{(y_i+1)/(x_i+1)\} - 1] . \end{aligned} \tag{16}$$

We calculate the $\operatorname{var}\left(\hat{O}_i \mid y_i\right)$ by conditioning on $\hat{T}_i$, as in (9). First, the variances of $\hat{O}_i$ and $\tilde{N}_i$, conditional on both y_i and $\hat{T}_i$, are equal

$$\operatorname{var}\left(\hat{O}_i \mid y_i, \hat{T}_i\right) = \operatorname{var}\left(\tilde{N}_i \mid y_i, \hat{T}_i\right).$$

Further, the expectation of $\hat{O}_i$, conditional on both y_i and $\hat{T}_i$, is equal to

$$E\left(\hat{O}_i \mid y_i, \hat{T}_i\right) = E\left(\tilde{N}_i \mid y_i, \hat{T}_i\right) - \hat{T}_i \approx N_i - \hat{T}_i ,$$

since $\tilde{N}_i$ is nearly unbiased for N_i . Using these results yields

$$\operatorname{var}\left(\hat{O}_i \mid y_i\right) = \operatorname{var}\left(\tilde{N}_i \mid y_i\right) + \operatorname{var}\left(\hat{T}_i \mid y_i\right). \tag{17}$$

An estimator of the variance of $\hat{O}_i$, conditional on y_i , is

$$\hat{v}\left(\hat{O}_i \mid y_i\right) = \hat{v}\left(\tilde{N}_i \mid y_i\right) + \hat{v}\left(\hat{T}_i \mid y_i\right) \tag{18}$$

where

$$\hat{v}\left(\hat{T}_i \mid y_i\right) = \hat{\pi}_i\left(1-\hat{\pi}_i\right) R_i^2 / r_i .$$

If we replace $\hat{v}\left(\hat{T}_i \mid y_i\right)$ in (18) with

$$\tilde{v}\left(\hat{T}_i \mid y_i\right) = \hat{\pi}_i\left(1-\hat{\pi}_i\right) R_i^2/(r_i-1) ,$$

which is an unbiased estimator of $\mathrm{var}\left(\hat{T}_i \mid y_i\right)$, then the

$$E\left\{\hat{v}\left(\hat{O}_i \mid y_i\right)\right\} \approx E\left\{\hat{v}\left(\tilde{N}_i \mid y_i\right)\right\} + \mathrm{var}\left(\hat{T}_i \mid y_i\right)$$

so that the bias of $\hat{v}\left(\hat{O}_i \mid y_i\right)$ is equal to the bias of $\hat{v}\left(\tilde{N}_i \mid y_i\right)$.

5.4. Simulation study

We performed a small simulation study to investigate the performance of the statistics, $\tilde{N}_i$ and $\hat{v}\left(\tilde{N}_i \mid y_i\right)$, in terms of the theoretical and simulated bias. Another objective was to assess the magnitude of $\mathrm{var}\left(\tilde{N}_i \mid y_i\right)$ for (R_i , r_i , y_i) configurations similar to those of the CUSSP survey. We calculated both bias and percentage bias (calculated as 100 × $\{\mu-E(\hat{\mu})\}/\mu$, where μ is a generic estimand); we present the latter to facilitate comparisons between different values of N_i. Details of the simulation technique appear in Appendix II. We used two values of (R_i, r_i, y_i), which corresponded to one sex and to one class of one sex from the CUSSP survey: (6539, 249, 134) corresponded to the total number of male undergraduates registered at Cornell, and (1589, 60, 35) corresponded to the number of sophomore males. We describe below the results of all simulations : we used 22 combinations of (N_i , π_i) for $R_i = 6539$ and 21 combinations for $R_i = 1589$. Table 1a gives the results of the simulations for several values of (N_i , π_i), using (6539, 249, 134); Table 1b gives the corresponding results using (1589, 60, 35).

The simulations showed that the bias of $\tilde{N}_i$ was negligible for $R_i = 6539$ with all

(N_i, π_i) combinations used, and was small (<10%) for $R_i = 1589$, except for combinations of large N_i with small π_i. For $R_i = 6539$, the coefficient of variation of $\widetilde{N}_i$, calculated as (standard deviation of $\widetilde{N}_i$) / E ($\widetilde{N}_i$), ranged from 1% to 30% (mean ± standard error: 13.7 ± 1.6) over all combinations of (N_i, π_i) examined, decreasing monotonically with increasing π_i for each N_i and increasing monotonically with N_i for each π_i. For $R_i = 1589$, the coefficient of variation of $\widetilde{N}_i$ ranged from 8% to 56% over the various combinations of N_i and π_i (27.8 ± 2.5), with a similar pattern. These patterns also hold for the theoretical and simulated standard deviations of $\widetilde{N}_i$. For $R_i = 1589$, the ratio of the simulated standard deviation to the theoretical standard deviation of $\widetilde{N}_i$ ranged from 0.94 to 1.82 (1.45 ± 0.06). For $R_i = 6539$, that ratio ranged from 1.02 to 2.39 (1.35 ± 0.08), with the three values greater than 2.0 only occurring for $\pi_i = 0.05$ with $N_i \geq 4000$. The latter were the only combinations tested using $R_i = 6539$ for which the percentage bias of $\widetilde{N}_i$ exceeded 0.06%. Simulations using small π_i produce values of $\widetilde{N}_i$ that are relatively small; hence, specification of large values of N_i in conjunction with small π_i results in appreciable bias of $\widetilde{N}_i$. In those circumstances, the discrepancy between the theoretical and simulated standard deviation of $\widetilde{N}_i$ can be substantial, and both standard deviations of $\widetilde{N}_i$ are relatively large. The combination of large N_i in conjunction with small π_i has the idiosyncrasy that the population at risk in the given time period is declared large but its members have a low probability of being sexually active during that period. Mathematically, the large theoretical standard deviation of $\widetilde{N}_i$ for this configuration arises because the first term in (12) is large, dominating the numerator of the ratio for the $E\left\{\operatorname{var}\left(\widetilde{N}_i \mid y_i, \hat{T}_i\right)\right\}$.

For a given value of N_i, the theoretical bias of $\hat{v}(\tilde{N}_i \mid y_i)$ decreases monotonically with increasing π_i. The bias of $\hat{v}(\tilde{N}_i \mid y_i)$ goes from negative to positive with increasing N_i for most π_i. The corresponding simulated bias, calculated as var$(\tilde{N}_i \mid y_i)$ -mean of $\hat{v}(\tilde{N}_i \mid y_i)$, generally followed the same pattern for $R_i = 1589$ but was always negative for $R_i = 6539$. For the latter case, in our simulations $\hat{v}(\tilde{N}_i \mid y_i)$ seemed consistently to provide an overestimate of the variance of $\tilde{N}_i$. For both configurations and every π_i examined, the correspondence between the theoretical bias of $\hat{v}(\tilde{N}_i \mid y_i)$ and its simulated counterpart was worst at the value of N_i for which the theoretical bias of $\hat{v}(\tilde{N}_i \mid y_i)$ was nearly zero (e.g., N_i =3000, π_i =0.05 in Table 1a).

6. EXTENSION TO STRATIFICATION WITHIN EACH SEX

We now consider two way stratification, in which we stratify the marked subpopulation (sexually active Cornell undergraduates) by both sex and college class. Not only are survey respondents stratified; respondents classify their marked partners as to sex and college class. The outsider subpopulation also is stratified by sex but not necessarily by college class. As before, our objective is to estimate the total size of the population of sex i at risk and the number of outsiders of sex i. Again, we must address the extra variation introduced into estimation of the number at risk by use of an estimate of the number of sexually active Cornell students.

6.1. Extended notation

Because a random sample was independently drawn from each stratum, the hypergeometric and approximating binomials models of Section 5.1 hold for each stratum. To account for the multiway stratification, however, we must introduce additional subscripts. As before, the subscript i denotes sex. The subscript j denotes the additional stratification variable with c levels (e.g., college class with 4 levels: freshman, sophomore, junior, senior). R_{ij}, T_{ij}, r_{ij}, t_{ij} and π_{ij} are as defined previously in Section 5.1, with the additional subcript j specifying college class. We denote the total number of Cornell undergraduates of sex i by $R_{i\cdot} = \sum R_{ij}$ and the number of those sexually active by $T_{i\cdot} = \sum T_{ij}$. We define the probability of an individual of sex i being sexually active as

$$\pi_{i\cdot} \equiv T_{i\cdot} / R_{i\cdot} = \sum_{j=1}^{c} R_{ij}\, \pi_{ij} / R_{i\cdot} \,,$$

which is a weighted average of the π_{ij}'s. For consistency of notation, we now denote the total size of the population of sex i at risk as $N_{i\cdot}$, and the number of outsiders of sex i is $O_{i\cdot} \equiv N_{i\cdot} - T_{i\cdot}$.

Parallel to the case of one way stratification, under the binomial model, the maximum likelihood estimators of T_{ij} and π_{ij} are $\widehat{T}_{ij} = t_{ij} R_{ij} / r_{ij}$ and $\widehat{\pi}_{ij} = t_{ij} / r_{ij}$, respectively. The corresponding estimators of $T_{i\cdot}$ and $\pi_{i\cdot}$ are $\widehat{T}_{i\cdot} = \sum \widehat{T}_{ij}$ and $\widehat{\pi}_{i\cdot} = \widehat{T}_{i\cdot} / R_{i\cdot}$, respectively, allowing each sex and class to have its own probability of being sexually active during the period in question. Consequently, $\widehat{T}_{ij}$ and $\widehat{T}_{i\cdot}$ are unbiased, consistent estimators of T_{ij} and $T_{i\cdot}$, respectively.

Let y_{ij} and $y_{i\cdot}$ be the total number of sexual contacts with individuals of sex i during the two month period reported by individuals of class j , and the grand total of sexual contacts with sex i reported by all surveyed Cornell students, respectively. All survey respondents classify their Cornell partners by college class. Therefore, we require three subscripts to tabulate Cornell sexual partners properly: i denotes the sex of the partner, j denotes the college class of the respondent and k denotes the college class of the partner. Let $x_{ij\cdot}$ be the total number of sexual contacts with Cornell undergraduates of sex i during the two month period reported by individuals of the opposite sex in class j. Let $x_{i\cdot k}$ denote the total number of sexual contacts with Cornell undergraduates of sex i and class k reported by all respondents of the opposite sex. We denote the corresponding number of sexual contacts with Cornell undergraduates of sex i, combined across all respondents' college classes, as $x_{i\cdot\cdot} = \sum x_{ij\cdot}$.

6.2. Estimators of the size of the total mixing population

A complication in extending estimation of the total number of individuals of sex i at risk ($N_{i\cdot}$) to the situation of two way stratification is the fact that the second stratification factor, college class, is irrelevant for those members of the unmarked group who are not college undergraduates. We approach the problem by extending the Bailey binomial model to a multinomial model and we derive an estimator of $N_{i\cdot}$. We also consider two variations of that model with their corresponding estimators. We discuss the strengths and weaknesses of the competing estimators.

We can extend Bailey's model given in (4) to be multinomial for each sex, with the population at risk consisting of five distinct subgroups, the four Cornell classes and the outsiders. Hence,

$$P\left(\underset{\sim}{X}_i = \underset{\sim}{x}_i \mid \underset{\sim}{T}_i, y_{i\cdot}\right) = \binom{y_{i\cdot}}{x_{i\cdot 1} x_{i\cdot 2} x_{i\cdot 3} x_{i\cdot 4}} \left(T_{i1} / N_{i\cdot}\right)^{x_{i\cdot 1}} \left(T_{i2} / N_{i\cdot}\right)^{x_{i\cdot 2}}$$

$$\times \left(T_{i3} / N_{i\cdot}\right)^{x_{i\cdot 3}} \left(T_{i4} / N_{i\cdot}\right)^{x_{i\cdot 4}} \left\{1 - \left(T_{i\cdot} / N_{i\cdot}\right)\right\}^{y_{i\cdot} - x_{i\cdot\cdot}}, \qquad (19)$$

where $x_{i\cdot k}$ is the total number of reported sexual contacts with Cornell students of sex i and class k and $y_{i\cdot}$-$x_{i\cdot\cdot}$ is the corresponding total number of contacts with outsiders of sex i. Under this model, the maximum likelihood estimator of the number of people of sex i at risk is

$$\tilde{N}_{i\cdot} = \hat{T}_{i\cdot} \left(y_{i\cdot} + 1\right) / \left(x_{i\cdot\cdot} + 1\right). \qquad (20)$$

This estimator has the advantages of allowing for different rates of sexual activity among Cornell undergraduates of different classes and allowing for the possibility of sampling with replacement with respect to sexual partners for all individuals surveyed, regardless of their college class. The var$\left(\tilde{N}_{i\cdot} \mid y_{i\cdot}\right)$, $\hat{v}\left(\tilde{N}_{i\cdot} \mid y_{i\cdot}\right)$, bias of $\tilde{N}_{i\cdot}$ and bias of $\hat{v}\left(\tilde{N}_{i\cdot} \mid y_{i\cdot}\right)$ are natural extensions of the formulae for one way stratification, substituting $\hat{T}_{i\cdot}$ and $\tilde{N}_{i\cdot}$ for $\hat{T}_i$ and $\tilde{N}_i$, respectively. For instance, the bias of $\tilde{N}_{i\cdot}$ is the product of the moment generating function of several binomial random variables (t_{ij} 's), which is no greater than the maximum of the biases of the individual t_{ij} 's.

We can impose a restriction on the model given in (19) by constraining the T_{ij} 's to reflect a common rate of sexual activity across all college classes for each sex. This is equivalent to assuming that for each i, the π_{ij} 's have a common value, say π_{i0}. Using this constraint is analogous to using the marginal totals in the analysis of multiway contingency tables by collapsing over the class of the respondent for the survey data and over class in the estimation of $T_{i\cdot}$. Under this assumption, an estimator of $N_{i\cdot}$ is

$$\tilde{N}^{*}_{i\cdot} = \tilde{T}_{i\cdot}\,(y_{i\cdot}+1)\,/\,(x_{i\cdot\cdot}+1)\,, \tag{21}$$

where $\tilde{T}_{i\cdot} = R_{i\cdot}(t_{i\cdot}/r_{i\cdot})$. The formulae for the var $\left(\tilde{N}^{*}_{i\cdot} \mid y_{i\cdot}\right)$, $\hat{v}\left(\tilde{N}^{*}_{i\cdot} \mid y_{i\cdot}\right)$, bias of $\tilde{N}^{*}_{i\cdot}$ and bias of $\hat{v}\left(\tilde{N}^{*}_{i\cdot} \mid y_{i\cdot}\right)$ correspond to those for one way stratification, with the appropriate substitutions made. The assumption of a common rate of sexual activity across college classes within each sex, is likely false. In fact, the CUSSP survey data indicate that the proportion of individuals sexually active increases with college class (i.e., number of years in a college program) for each sex (see Section 7). Thus, we do not favor this estimator for the CUSSP survey data.

A third alternative is to consider individuals of each college class as belonging to disjoint populations at risk. This corresponds to a product of four independent binomial models for each sex. Hence, if we seek to estimate the combined number at risk across the four disjoint populations ($N_{i\cdot}$), we sum estimates that are made separately, each using survey data only from respondents in a given college class. Thus, the combined estimator is

$$\hat{N}_{i\cdot} = \sum_{j=1}^{c} \tilde{N}_{ij} = \sum_{j=1}^{c} \hat{T}_{ij}\,(y_{ij}+1)\,/\,(x_{ij\cdot}+1)\,. \tag{22}$$

This model implies that there is no sexual contact between members of different college classes. This is certainly not the case for Cornell students; therefore, we should not apply this estimator to the CUSSP data. We discuss this estimator, however, because it is apt to have an initial attraction, when the objective is to estimate the population subtotal in stratum i for a population having multiway stratification. For situations in which this model is appropriate, we can find the variance and bias of $\hat{N}_{i\cdot}$ by summing those of the individual $\hat{N}_{ij}$'s. We illustrate this estimator in the next section using CUSSP survey data, but we do not calculate its variance since the model does not apply to the CUSSP survey.

7. SIZE OF THE TOTAL MIXING POPULATION FROM THE CUSSP SURVEY

In this section, we illustrate the methodology for one way and two way stratification, using data from the CUSSP survey. Table 2 gives information required to estimate the number of sexually active Cornell undergraduates of each sex and college class for the Fall of 1989 : the enrollment of each sex in each class (R_{ij}), the number of each sex and class that responded to the direct version of the CUSSP survey (r_{ij}) and the number responding that met the survey's definition of sexually active (t_{ij}). For each class, the proportion of sexually active male respondents was much lower than that for females, although, for both sexes, the proportion of sexually active individuals increases with college class. Table 3 gives a summary of the number of sexual contacts with members of the opposite sex ($x_{ij\cdot}$ and y_{ij}) reported by respondents of each sex and college class. For each class, the total number of female partners reported by male respondents was lower than the corresponding number of male partners reported by females.

Table 4 gives separate estimates of the number of sex i at risk due to contact with students of the opposite sex in class j ($\hat{N}_{ij}$) and their standard deviations; the three competing estimates of $N_{i\cdot}$ also appear. For each sex, $\hat{N}_{i\cdot}$ is the largest of the three estimates, $\tilde{N}^{*}_{i\cdot}$ is smallest and $\tilde{N}_{i\cdot}$ is intermediate. For example, the estimates of the size of the female population who have sexual contact with Cornell undergraduate males are: $\widehat{N}_{f\cdot} = 4186$, $\widetilde{N}^{*}_{f\cdot} = 3929$ and $\widetilde{N}_{f\cdot} = 3993$. That $\widetilde{N}_{i\cdot}$ is intermediate seems consistent with the hybrid nature of the estimator: it allows each class to have its own π_{ij} while using marginal totals for the number of sexual contacts.

If all Cornell undergraduates were sexually active, then $R_{ij} = T_{ij}$. Under this assumption, the stratification is of no consequence since the number of sexually active people is known. Thus, Bailey's estimators of N_i and its variance, (5) and (6), are appropriate. The total number of male undergraduates registered at Cornell in the Fall of 1989, R_m, was 6539; the corresponding number of females, R_f, was 5211. We find that $\widehat{N}_m = 12982$, with a standard deviation of 1101, and $\widehat{N}_f = 8956$, with a standard deviation of 1008. As expected, these total population estimates are considerably greater than all corresponding combined estimates and they can serve as rough upper bounds for the population sizes of each sex. For both sexes, the estimated coefficient of variation of $\widetilde{N}^{*}_{i\cdot}$, $\widetilde{N}_{i\cdot}$ and the Bailey estimator were close (females: 10.8, 10.8 and 11.3 %, respectively; males: 7.5, 7.6, and 8.5 %, respectively).

8. CONCLUSIONS

In this paper, we have formulated a mark-recapture model, for use with survey data in which members of the targeted group give information about sexual contacts with outsiders so as to estimate the size of the total population that has direct sexual contact with a marked group. Consequently, we can estimate the number of outsiders who have sexual contact with members of the marked group. We have derived an estimator of the variance of the estimated number at risk, which takes into account estimation of the size of the marked population, along with an expression for the approximate bias of that variance estimator. We have provided point estimators of the number of outsiders and its variance. The estimates of the total population and the size of the outsider subgroup provide initial parameter estimates for mathematical models of disease spread via social and sexual mixing. The variance estimates for these parameters provide a guide for the range of configurations to use in simulation studies of the mixing models.

ACKNOWLEDGEMENTS

The authors thank Drs. Charles E. McCulloch and Douglas S. Robson, anonymous reviewers and the editor for their suggestions that greatly improved this manuscript. C. Castillo-Chavez acknowledges the support of grants from NIAID and the USDA Hatch program for part of this work.

APPENDIX I: Expected value of the estimator of the variance

In this appendix we give the formulae required to calculate the approximate expected value of $\hat{v}\left(\tilde{N}_i \mid y_i\right)$. Using (14) we can write

$$E\left\{\hat{v}\left(\tilde{N}_i \mid y_i\right)\right\} = E\left[E\left\{\hat{v}\left(\tilde{N}_i \mid y_i, \hat{\pi}_i, x_i\right)\right\}\right]$$

$$= E\left[E\left\{A\left(\hat{\pi}_i\right)/C\left(\hat{\pi}_i\right) \mid y_i, \hat{\pi}_i, x_i\right\}\right] + E\left[E\left\{\tilde{N}_i^2\left(\left[\left\{B\left(\tilde{N}_i/2\right)\right\}^{r_i} - \left\{B\left(\tilde{N}_i\right)\right\}^{2r_i}\right] \mid y_i, \hat{\pi}_i, x_i\right)\right\}\right]. \quad \text{(i)}$$

The second term of the RHS of (i) contributes almost nothing to the expectation we seek. Using a zero order Taylor series expansion, the first term of the RHS of (i) is approximately equal to

$$E\big[E\big(A(\hat{\pi}_i) \mid y_i, \hat{\pi}_i, x_i\big)\big] / E\big[E\big(C(\hat{\pi}_i) \mid y_i, \hat{\pi}_i, x_i\big)\big], \qquad \text{(ii)}$$

Both the numerator and the denominator of (ii) are 7th order polynomials in π_i.

Denoting the kth descending factorial of r_i and y_i as $r_i^{(k)}$ and $y_i^{(k)}$, respectively,

$$\begin{aligned}
&E\big[E\big(A(\hat{\pi}_i) \mid y_i, \hat{\pi}_i, x_i\big)\big] \\
&\quad = y_i\,(y_i+1)\big(R_i^2 / r_i^2\big)\Big[\big\{\pi_i r_i + \pi_i^2 r_i^{(2)}\big\}\big(L_0 / r_i^2\big) + \big\{\pi_i r_i + 3\pi_i^2 r_i^{(2)} + \pi_i^3 r_i^{(3)}\big\}\big(L_1 / r_i^3\big) \\
&\qquad + \big\{\pi_i r_i + 7\pi_i^2 r_i^{(2)} + 6\pi_i^3 r_i^{(3)} + \pi_i^4 r_i^{(4)}\big\}\big(L_2 / r_i^4\big) \\
&\qquad + \big\{\pi_i r_i + 15\pi_i^2 r_i^{(2)} + 25\pi_i^3 r_i^{(3)} + 10\pi_i^4 r_i^{(4)} + \pi_i^5 r_i^{(5)}\big\}\big(L_3 / r_i^5\big)\Big]
\end{aligned}$$

where

$$\begin{aligned}
L_0 &= -(r_i y_i + r_i)^{-1}\big\{1 + 3y_i(\pi_i R_i / N_i) + (\pi_i R_i / N_i)^2 y_i^{(2)}\big\}, \\
L_1 &= 1 + y_i(\pi_i R_i / N_i) - \{7\,(r_i - 1) / (r_i y_i + r_i)\}\big\{1 + 3y_i(\pi_i R_i / N_i) + (\pi_i R_i / N_i)^2 y_i^{(2)}\big\}, \\
L_2 &= 3\,(r_i - 1)\{1 + y_i(\pi_i R_i / N_i)\} - \{2\,(r_i - 2) / (r_i y_i + r_i)\}\big\{1 + 3y_i(\pi_i R_i / N_i) + (\pi_i R_i / N_i)^2 y_i^{(2)}\big\}, \\
L_3 &= \big(r_i^{(3)} / r_i\big)\,\{1 + y_i(\pi_i R_i / N_i)\} - \{(r_i - 3) / (r_i y_i + r_i)\}\big\{1 + 3y_i(\pi_i R_i / N_i) + (\pi_i R_i / N_i)^2 y_i^{(2)}\big\}.
\end{aligned}$$

Likewise,

$$\begin{aligned}
&E\big[E\big(C(\hat{\pi}_i) \mid y_i, \hat{\pi}_i, x_i\big)\big] \\
&\quad = \big(K_0 / r_i^3\big)\big\{\pi_i r_i + 3\pi_i^2 r_i^{(2)} + \pi_i^3 r_i^{(3)}\big\} + \big(K_1 / r_i^2\big)\big\{\pi_i r_i + \pi_i^2 r_i^{(2)}\big\} + K_2 \pi_i + K_3,
\end{aligned}$$

where

$$\begin{aligned}
K_0 &= 1 + \{4y_i / (y_i+1)\}(1 + y_i \pi_i R_i / N_i) + \big\{7y_i^2 r_i^{(2)} / (r_i y_i + r_i)^2\big\}\big\{1 + 3y_i(\pi_i R_i / N_i) + (\pi_i R_i / N_i)^2 y_i^{(2)}\big\} \\
&\quad + \big\{4y_i^3 r_i^{(3)} / (r_i y_i + r_i)^3\big\}\,\big\{1 + 7y_i(\pi_i R_i / N_i) + 6(\pi_i R_i / N_i)^2 y_i^{(2)} + (\pi_i R_i / N_i)^3 y_i^{(3)}\big\} \\
&\quad + \big\{1 + 15y_i(\pi_i R_i / N_i) + 25(\pi_i R_i / N_i)^2 y_i^{(2)} + 10(\pi_i R_i / N_i)^3 y_i^{(3)} + (\pi_i R_i / N_i)^4 y_i^{(4)}\big\}\, y_i^4 r_i^{(4)} / (r_i y_i + r_i)^4, \\
K_1 &= \big\{6y_i^2 r_i / (r_i y_i + r_i)^2\big\}\Big[1 + 3y_i(\pi_i R_i / N_i) + (\pi_i R_i / N_i)^2 y_i^{(2)} + \{2y_i(r_i - 1) / (r_i y_i + r_i)\} \\
&\qquad \times \big\{1 + 7y_i(\pi_i R_i / N_i) + 6(\pi_i R_i / N_i)^2 y_i^{(2)} + (\pi_i R_i / N_i)^3 y_i^{(3)}\big\} \\
&\quad + \big\{1 + 15y_i(\pi_i R_i / N_i) + 25(\pi_i R_i / N_i)^2 y_i^{(2)} + 10(\pi_i R_i / N_i)^3 y_i^{(3)} + (\pi_i R_i / N_i)^4 y_i^{(4)}\big\} y_i^2 r_i^{(3)} / \big(r_i^3 (y_i + 1)^2\big)\Big],
\end{aligned}$$

$$K_2=\left\{1+7y_i(\pi_iR_i/N_i)+6(\pi_iR_i/N_i)^2y_i^{(2)}+(\pi_iR_i/N_i)^3y_i^{(3)}\right\}4y_i^3r_i/(r_iy_i+r_i)^3$$

$$+\left\{1+15y_i(\pi_iR_i/N_i)+25(\pi_iR_i/N_i)^2y_i^{(2)}+10(\pi_iR_i/N_i)^3y_i^{(3)}+(\pi_iR_i/N_i)^4y_i^{(4)}\right\}6y_i^4r_i^{(2)}/(r_iy_i+r_i)^4,$$

$$K_3=\left\{1+15y_i(\pi_iR_i/N_i)+25(\pi_iR_i/N_i)^2y_i^{(2)}+10(\pi_iR_i/N_i)^3y_i^{(3)}+(\pi_iR_i/N_i)^4y_i^{(4)}\right\}y_i^4r_i/(r_iy_i+r_i)^4.$$

APPENDIX II: Simulation method

We programmed all simulations in the DATA step of SAS[15], using 1000 replications per simulation. For each combination of (R_i, r_i, y_i) with (N_i, π_i), we calculated $E(\tilde{N}_i|y_i)$, $\mathrm{var}(\tilde{N}_i \mid y_i)$, $E\{\hat{v}(\tilde{N}_i \mid y_i)\}$ and bias of $\hat{v}(\tilde{N}_i \mid y_i)$. For each replication, we drew the variate t_i randomly from a binomial distribution with parameters (r_i, π_i), and the variate x_i randomly from a binomial distribution with parameters $(y_i, R_i\hat{\pi}_i/N_i)$; we calculated the statistics, $\tilde{N}_i$ and $\hat{v}(\tilde{N}_i \mid y_i)$. We generated all binomial variates using the built-in binomial random number generator, with the time on the computer clock as the seed. We calculated the sample mean and variance of $\tilde{N}_i$ and $\hat{v}(\tilde{N}_i \mid y_i)$ from the 1000 replications, as well as the simulated bias of $\tilde{N}_i$ and $\hat{v}(\tilde{N}_i \mid y_i)$.

We used two values of (R_i, r_i, y_i), which correspond to one sex and to one sex-class stratum from the CUSSP survey; (6539, 249, 134) corresponded to total number of male undergraduates registered at Cornell and (1589, 60, 35) corresponded to the number of sophomore males. For each (R_i, r_i, y_i) configuration, we used $\pi_i = 0.05, 0.10, 0.25, 0.50, 0.75$ in combination with several values of N_i. For $R_i = 6539$ and $\pi_i \le 0.25$, $N_i = 700, 1500, 2000, 3000, 4000, 6500, 9000$; for $\pi_i > 0.25$, $N_i = 2000, 3000, 4000, 6500, 9000$. For $R_i = 1589$ and $\pi_i < 0.25$, $N_i = 150, 400, 700, 2000, 4000$; for $\pi_i \ge 0.25$, $N_i = 700, 1500, 2000, 4000,$

9000.

REFERENCES

1. De Gruttola, V. and Lagakos, S. 'Epidemic models, empirical studies, and uncertainty', in *Mathematical and Statistical Approaches to AIDS Epidemiology*, C. Castillo-Chavez (Ed.), Lecture Notes in Biomathematics **83**, 38-57, Springer-Verlag, Berlin, 1989.
2. Sattenspiel, L. and Castillo-Chavez, C. 'Environmental context, social interactions, and the spread of HIV', *Am. J. Human Biol.*, **2**, 397-417 (1990).
3. Buesenberg, S. and Castillo-Chavez, C. 'Interaction, pair formation and force of infection terms in sexually transmitted diseases', in *Mathematical and Statistical Approaches to AIDS Epidemiology*, C. Castillo-Chavez (Ed.), Lecture Notes in Biomathematics **83**, 289-300, Springer-Verlag, Berlin, 1989.
4. Buesenberg, S. and Castillo-Chavez, C. 'A general solution of the problem of mixing subpopulations and its application to risk- and age-structured epidemic models for the spread of AIDS', *IMA Journal of Mathematics Applied in Medicine and Biology*, **8**, 1-29 (1991).
5. Stigum, H., Grønnesby, J.K., Magnus, P., Sundet,.J.M. and Bakketeig, L.S. 'The potential for spread of HIV in the heterosexual population in Norway: a model study', *Statistics in Medicine*, **10**, 1003-1023 (1991).
6. Bailey, N.T.J. 'On estimating the size of mobile populations from capture-recapture data', *Biometrika*, **38**, 293-306 (1951).
7. Crawford, C.M., Schwager, S.J. and Castillo-Chavez, C. 'A methodology for

asking sensitive questions among college undergraduates', Technical Report # BU-1105-M in the Biometrics Unit Series, Cornell University, Ithaca, NY (1990).

8. Wittes, J.T. 'Applications of a multinomial capture-recapture model to epidemiological data', *J. Am. Statist. Assoc.*, **69**, 93-97 (1974).

9. Goldberg, J.D. and Wittes, J.T. 'The estimation of false negatives in medical screening', *Biometrics*, **34**, 77-86 (1978).

10. Fay, R.E., Turner, C.F., Klassing, A.D. and Gagnon, J. 'Prevalence and patterns of same gender sexual contact among men', *Science*, **243**, 338-348 (1989).

11. Seber, G.A.F. *The Estimation of Animal Abundance and Related Parameters (second edition)*, Charles Griffin & Company Ltd., London, 1982.

12. Blythe, S.P., Castillo-Chavez, C. and Casella, G. 'Empirical methods for the estimation of the mixing probabilities for socially-structured populations from a single survey sample', *Mathematical Population Studies* (in press, 1991).

13. Pollack, K.H., Nichols, J.D., Brownie, C. and Hines, J.E. 'Statistical Inference for Capture-Recapture Experiments', *Wildlife Monographs*, **107** [*J. Wildl. Manag.*, Suppl. **54**] (1990).

14. Cone, R.S., Robson, D.S, and Krueger, C.C. 'Failure of statistical tests to detect assumption violations in the mark-recapture population estimation of brook trout in Adirondack ponds', *North Am. J. Fisheries Management*, **8**, 489-496 (1988).

15. SAS Institute Inc. Version 5 Edition, SAS Institute, Inc., Cary, NC, 1985.

Table 1a. Simulation results for several values of π_i and N_i, using $(R_i, r_i, y_i) = (5539, 249, 134)$. The standard deviation (SD) and simulated SD of $\tilde{N}_i$, the theoretical and simulated biases of $\tilde{N}_i$ and $\hat{v}(\tilde{N}_i \mid y_i)$, and the coefficient of variation (CV) of $\tilde{N}_i$ (%) are given. All biases are presented as percentage bias, calculated as $100 \times \{\mu - E(\hat{\mu})\}/\mu$, where μ is a generic estimand.

			N_i		
π_i =0.05	700	2000	3000	4000	9000
SD of $\tilde{N}_i$	52.0	325.4	599.4	906.5	2653.0
sim. SD	75.2	442.4	862.5	1918.9	6337.7
CV of $\tilde{N}_i$	7.4	16.3	20.0	22.7	30.0
bias of $\tilde{N}_i$	0.00	0.00	0.02	0.06	1.72
sim bias	-0.23	-0.36	0.13	-1.48	-3.91
bias of $\hat{v}(\tilde{N}_i \mid y_i)$	-48.13	-14.80	0.58	12.12	43.13
sim. bias	-22.83	-22.52	-27.49	-51.51	-45.77
π_i =0.25					
SD of $\tilde{N}_i$	-	72.9	223.5	392.4	1504.6
sim. SD	-	85.3	242.5	409.1	1850.8
CV of $\tilde{N}_i$	-	3.6	7.4	9.8	16.7
bias of $\tilde{N}_i$	-	0.00	0.00	0.00	0.00
sim bias	-	-0.05	-0.13	0.06	-0.28
bias of $\hat{v}(\tilde{N}_i \mid y_i)$	-	-6.13	-1.46	2.32	17.38
sim. bias	-	-12.17	-7.33	-5.40	-16.03
π_i =0.50					
SD of $\tilde{N}_i$	-	-	-	156.3	986.7
sim. SD	-	-	-	172.5	1007.0
CV of $\tilde{N}_i$	-	-	-	3.9	11.0
bias of $\tilde{N}_i$	-	-	-	0.00	0.00
sim bias	-	-	-	0.05	0.47
bias of $\hat{v}(\tilde{N}_i \mid y_i)$	-	-	-	0.10	9.10
sim. bias		-	-	-6.31	-3.99

Table 1b. Simulation results for several values of π_i and N_i, using (R_i, r_i, y_i) = (1589, 60, 35).

	N_i			
π_i =0.05	400	700	2000	4000
SD of $\widetilde{N}_i$	116.7	239.2	761.4	*
sim. SD	189.8	319.9	903.8	*
CV of $\widetilde{N}_i$	31.1	38.2	56.4	*
bias of $\widetilde{N}_i$	6.28	10.61	32.53	*
sim bias	0.32	4.30	25.24	*
bias of $\hat{v}(\widetilde{N}_i \mid y_i)$	14.29	31.48	65.24	*
sim. bias	22.70	29.22	56.40	*
π_i =0.25				
SD of $\widetilde{N}_i$	-	82.4	504.1	1128.1
sim. SD	-	118.0	806.3	2047.9
CV of $\widetilde{N}_i$	-	11.8	25.3	32.0
bias of $\widetilde{N}_i$	-	0.00	0.29	4.14
sim bias	-	0.12	-1.08	4.76
bias of $\hat{v}(\widetilde{N}_i \mid y_i)$	-	-11.17	33.37	56.11
sim. bias	-	-27.93	-33.32	-4.05
π_i =0.50				
SD of $\widetilde{N}_i$	-	-	359.9	1042.5
sim. SD	-	-	448.0	1779.9
CV of $\widetilde{N}_i$	-	-	18.0	26.0
bias of $\widetilde{N}_i$	-	-	0.00	0.14
sim bias	-	-	-0.26	-1.66
bias of $\hat{v}(\widetilde{N}_i \mid y_i)$	-	-	21.78	44.32
sim. bias	-	-	-19.69	-38.14

* N_i is too large for the π_i specified (i.e., led to underflow problems)

Table 2. Cornell undergraduates enrolled in the Fall of 1989 (R_{ij}), number that responded to the direct version of the CUSSP survey (r_{ij}) and the number of respondents that were sexually active (t_{ij}).

Females

	Freshman	Sophomore	Junior	Senior	Total
R_{fj}	1278	1308	1277	1348	5211
r_{fj}	68	68	61	56	253
t_{fj}	21	26	36	28	111

Males

	Freshman	Sophomore	Junior	Senior	Total
R_{mj}	1673	1589	1591	1686	6539
r_{mj}	79	60	63	47	249
t_{mj}	5	4	6	7	22

Table 3. The total number of sexual contacts with the opposite sex (combined over participants from each class) reported by undergraduate respondents to the direct version of the CUSSP survey. Each row corresponds to the class of the respondents.

Female respondents' male partners

	Cornell undergraduate ($x_{mj\cdot}$)	Other	Total (y_{mj})
Freshman	15	14	29
Sophomore	21	14	35
Junior	23	18	41
Senior	8	21	29

Male respondents' female partners

	Cornell undergraduate ($x_{fj\cdot}$)	Other	Total (y_{fj})
Freshman	12	4	16
Sophomore	5	7	12
Junior	5	6	11
Senior	9	6	15

Table 4. The estimated size of the population of sex i having sexual contact with Cornell undergraduates. Individual estimates are given for each stratum (standard deviation given in parentheses) as well as estimates for each sex of the three competing combined population size estimators. The estimated number of sexually active Cornell undergraduates in each class ($\hat{T}_{ij}$) and the estimated total for each sex $\hat{T}_{i\cdot}$ also are given.

			$\tilde{N}_{ij}$				
i	Freshman	Sophomore	Junior	Senior	$\hat{N}_{i\cdot}$	$\tilde{N}^{*}_{i\cdot}$	$\tilde{N}_{i\cdot}$
Female	517	1083	1508	1078	4186	3929	3993
	(52)	(249)	(344)	(176)	-	(426)	(433)
Male	199	173	264	837	1473	1148	1219
	(18)	(5)	(20)	(168)	-	(86)	(92)

			$\hat{T}_{ij}$		$\hat{T}_{i\cdot}$
Female	395	500	754	674	2323
Male	106	106	152	251	615