

Statistical Methods for the Analysis of Time–Location Sampling Data

John M. Karon and Cyprian Wejnert

ABSTRACT *Time–location sampling (TLS) is useful for collecting information on a hard-to-reach population (such as men who have sex with men [MSM]) by sampling locations where persons of interest can be found, and then sampling those who attend. These studies have typically been analyzed as a simple random sample (SRS) from the population of interest. If this population is the source population, as we assume here, such an analysis is likely to be biased, because it ignores possible associations between outcomes of interest and frequency of attendance at the locations sampled, and is likely to underestimate the uncertainty in the estimates, as a result of ignoring both the clustering within locations and the variation in the probability of sampling among members of the population who attend sampling locations. We propose that TLS data be analyzed as a two-stage sample survey using a simple weighting procedure based on the inverse of the approximate probability that a person was sampled and using sample survey analysis software to estimate the standard errors of estimates (to account for the effects of clustering within the first stage [locations] and variation in the weights). We use data from the Young Men’s Survey Phase II, a study of MSM, to show that, compared with an analysis assuming a SRS, weighting can affect point prevalence estimates and estimates of associations and that weighting and clustering can substantially increase estimates of standard errors. We describe data on location attendance that would yield improved estimates of weights. We comment on the advantages and disadvantages of TLS and respondent-driven sampling.*

KEYWORDS *Time–location sampling, HIV, Statistical methods*

INTRODUCTION

Public health scientists use time–location sampling (TLS; also known as time-space sampling) to collect data from hard-to-reach populations, such as men who have sex with men (MSM), who can be found at identifiable locations. For example, the Centers for Disease Control and Prevention (CDC) used TLS in the Young Men’s Survey (YMS) to enroll 3,492 MSM ages 15–22 years in seven large cities during 1994–1998.¹ CDC recruited these men from a variety of types of venues (including dance clubs, bars, businesses, parks, and street locations) where MSM could be found. CDC collected public health data, including behavioral data and HIV and

Karon was with the Emergent Corporation, 2505 Elfego Road NW, Louisville, KY, USA; Wejnert is with the Division of HIV/AIDS Prevention, Centers for Disease Control and Prevention, Atlanta, GA, USA.

Correspondence: John M. Karon, 2505 Elfego Road NW, Albuquerque, NM 87107-3010, USA. (E-mail: jkaron@earthlink.net)

The findings and conclusions presented here are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

hepatitis B status. The prevalence of HIV at such a young age was high, 5.6% and 8.6% among men ages 15–19 and 20–22 years, respectively.

YMS investigators collected data using standard TLS methods.² First, formative research attempted to identify all venues (locations) at which enough persons from the population of interest could be found on some day during a period of at least 4 hours to make recruitment worthwhile. The potential sampling periods for each day were defined; there could be more than one period on a day at a venue (see MacKellar et al.³ for a hypothetical example). These venues and sampling times formed the sampling frame from which locations were selected; the sampling frame could be redefined monthly based on updated attendance information (venues could be added or deleted). Next, investigators chose a random sample of venues (a simple random sample without replacement) of a specified size from this sampling frame. For each venue selected, the investigators chose a day and then a time period for sampling at each venue; if there was more than one potential sampling period, the period during which sampling took place was chosen at random or considering administrative constraints. Finally, the investigators chose a sample of persons entering the venue from among those appearing to meet the eligibility criteria and counted the number of attendees appearing to meet these criteria. Investigators used procedures to avoid enrolling a person more than once.

The published analyses of YMS regarded the data as if they were obtained from a simple random sample. For example, the investigators estimated prevalence as the number with the condition, divided by the number with data; they used the binomial distribution to obtain confidence intervals and logistic regression to evaluate the association between an outcome and a risk factor, adjusted for covariates. Point estimates and confidence intervals from such an analysis describe a population resembling those actually sampled (over-representing persons who visit these locations more frequently) but not the population from which those sampled was drawn unless the persons studied were a simple random sample from an infinite population and data from pairs of persons were independent. Because the probability of sampling depends on the frequency with which a person attends venues, the probability of accepting an invitation to participate, and the sampling fractions at the venues he attends, the simple random sample assumption is false. The independence assumption is false if there is sufficient variation in the mean value of an outcome among the locations. Thus, estimates of prevalence from such an analysis may be biased if outcomes are associated with frequency of attendance at locations; estimates of associations between outcomes and personal characteristics may be biased if these characteristics are associated with frequency of attendance. Estimates of standard errors are too small if outcomes or characteristics cluster within locations or if there is differential probability of enrollment in the study; we will account for the latter by proposing a weighted analysis.

We assume that the population of interest is the population of persons attending locations, rather than a population resembling those actually sampled. In this case, a more appropriate analysis of a TLS study is that of a multi-stage cluster sample survey. The primary sampling units (PSUs) are the locations (the first stage). The sampling of persons within the PSUs is a second stage. The probability of being sampled depends on many factors, including whether a person goes to any venue, whether he goes to the venue at which sampling takes place during the sampling period, and the proportion of eligible persons sampled. We show how to define weights which approximately reflect the differential probability of enrollment in the study; we include the effect of all factors that affect the sampling probability in these

weights, so that the data can be analyzed as a two-stage rather than a three-stage sample (software for analyzing a two-stage sample is more readily available than for a three-stage sample). Using these weights and the location from which a person was enrolled (the cluster) in sample survey software accounts for the effects of variation among the weights and of clustering on the standard errors of estimates.

After summarizing the relevant sample survey theory, we propose a method to compute approximate weights for TLS data. We illustrate these methods by analyzing data from four of the six metropolitan areas in which CDC conducted a second TLS study of MSM, YMS Phase II, which enrolled 2,942 MSM ages 23–29 years during 1998–2000.⁴ The analyses demonstrate the effects of failing to use weights and accounting for clustering to estimate the prevalences and the precision of these estimates for HIV, hepatitis B, and unprotected anal intercourse (UAI) and to evaluate the associations between UAI and the other two outcomes. In the “Discussion” section, we comment on data collection for time–location surveys and the analysis of these data, and on the use of other data collection methods.

WEIGHTING AND VARIANCE ESTIMATION IN A SAMPLE SURVEY

We review standard sample survey analysis methods which account for among-person variation in sampling probabilities through the definitions of weights. We review the design effect, which quantifies the increase in estimates of standard errors as a result of variation in these weights and of clustering.

In a sample survey analysis, a weight w_i , the inverse of the probability of being sampled, is defined for each person. Let x_i and w_i be the observed quantity of interest and the weight, respectively, for the i th person in the sample. Then the estimated mean value in the sample is

$$\bar{x} = \Sigma w_i x_i / \Sigma w_i \quad (1)$$

Note that the mean is unchanged if all the weights are multiplied by a constant. The variance of an estimate based on a sample survey is a sum of the variances for each PSU. Each of these variances is an expression times $1 - \text{fpc}$, where fpc (the finite population correction factor) is the probability that the PSU was sampled. Without this correction, the estimated variance is too large if a substantial proportion of the PSUs are sampled.

In a sample with unequal probabilities of sampling or clustering, the variance of an estimate is likely to be larger than the variance obtained from a simple random sample of the same size. This increase is quantified as the *design effect* (DE), defined as the variance computed based on the sampling design, divided by the variance that would be obtained if the sample were a simple random sample. For example, if we estimate a prevalence with an observed value of p from a sample of size n and ignore the finite population correction factor, the denominator of the DE is $p(1-p)/n$. If all weights are equal, the design effect is approximately $1 + (\bar{n} - 1)\rho$ where \bar{n} is the mean cluster size and ρ is the *intracluster* (or *intra-class*) *correlation coefficient* (ICC; see Kish.)^{5(p258)} The ICC measures the variability among clusters compared to the variability in the entire population. The ICC is 1 if x is constant within each cluster. If the proportion with $x=1$ is constant across clusters, then the ICC is negative and is equal to $-1/(n-1)$ if all clusters are of size n (this is the minimum value for the ICC). Note that the ICC may be different for different outcomes of

interest. If the ICC is positive, it is plausible that the DE should increase as the mean cluster size or ICC increases, because there is more effect of the clusters sampled on the point estimate. Zou and Donner⁶ give three estimators (one based on analysis of variance and one based on Pearson's correlation coefficient, and a kappa-type estimator) of the ICC and their variances for binary data x under the hypotheses that the probability $\Pr(x=1)$ is constant across clusters, and that observations from different clusters are independent.

With unequal weights, it is plausible that the DE should increase as the variation among the weights increases, because the particular sample of persons selected for the study has more effect on an estimate. Kish⁵ quantified the increase in the design effect for a study with weights that are constant within strata but vary among strata (Eq. 11.7.7). This expression is 1 plus a sum of terms, one term for each pair of strata (for a TLS survey, locations). For two strata with sample sizes n_1 and n_2 and weights w_1 and w_2 , with mean weight \bar{w} , this term is $\bar{w}^2 CV^2 / (2N^2 w_1 w_2)$, where CV is the coefficient of variation of the weights in these two strata (the standard deviation of the weights divided by the mean of the weights) and N is the total sample size in the study. Thus, in a TLS study, the design effect likely increases with the square of the coefficient of variation of the weights, as well as with the ICC and cluster size.

DEFINITION OF WEIGHTS FOR A TIME-LOCATION SAMPLE STUDY

The prevalence of a sexually transmitted infection, such as HIV or hepatitis B, in MSM may be associated with the frequency of attendance at venues in the time-location study (hereafter also abbreviated as TLS) sampling frame. As noted in the "Introduction" section, if this were true, an unweighted estimate of prevalence would be biased compared to the prevalence in the population of MSM attending the venues at which sampling took place. In this section, we suggest a simple expression as an approximation to the probability with which a person is sampled for a TLS.

The sampling probability p for a person enrolled in a TLS is similar to the cumulative probability from a negative binomial distribution, as shown by Eq. 2 below. Let s_i be the probability that the person is enrolled during the i th sampling event, given the person's behavior (how often the person attends locations in the sampling frame, which location the person attends, and when). Then

$$p = s_1 + (1 - s_1)s_2 + (1 - s_1)(1 - s_2)s_3 + \dots \quad (2)$$

That is, the person could be enrolled during the first sampling event, or not enrolled during that event but enrolled during the second, or not enrolled during the first two events but enrolled during the third, etc. If the s_i are constant, the summands of Eq. 2 are the successive terms of the negative binomial probability distribution.

CDC investigators attempted to estimate p in analyzing the first YMS from interview information on how often a man went to the venues in the sampling frame and the probability distribution of the types of venues he attended. The probability that a man was sampled during the i th sampling event is the product of four probabilities: γ_i , the probability that a man attends any venue in the sampling frame on the day when that sampling event takes place; α_i , the probability that he attends

the venue at which sampling takes place given that he attends some venue on that day; β_i , the proportion of potential venue sampling time on that day during which sampling took place (e.g., β_i is 1 if there was only one potential sampling period and men were recruited during the entire period, and is 2/3 if sampling took place during an entire 4-h period but there was also a potential 2-h period); and f_i , the sampling fraction at that sampling event. Thus

$$s_i = \gamma_i \alpha_i \beta_i f_i \quad (3)$$

If f_i is computed as the number who participate divided by the number believed eligible, and if we do not stratify the response probability and the sampling fraction by personal characteristics, then f_i includes the effect of nonresponse.

MacKellar et al.² estimated s_i for each man under the assumptions that attendance at all venues of a particular type was equally probable ($\alpha_i=1$) and that for all sampling events, β_i was 1. However, the resulting numerical estimates of the sampling probability p were approximately 1 for nearly all men; this implies that nearly every man in the population was sampled, which clearly was not true (sampling fractions were less than 0.25 even at the end of the study). There were a large number of sampling events in each metropolitan area (1,592 sampling events in seven areas).¹ Therefore, for each man sampled, Eq. 2 is a sum of a very large number of terms, each of which may be small. Such a sum is very difficult to compute accurately, which likely led to the failure of this procedure.

Therefore, we use an alternative approach. To obtain a simple approximation to a person's sampling probability, we assume that, for a person in the sample, s_i is constant and equal to s (s may vary among persons). This will be true if, for that person, γ_i , β_i , and the α_i are constant during the study period (i.e., the person's venue attendance pattern is unchanged during the study), and the sampling fractions are equal at all venues the person attended. Assume that there are k sampling events. Since the probability of not being sampled at each event is $1-s$, the probability of being sampled is

$$p = 1 - (1-s)^k = ks - k(k-1)s^2/2 + \dots \quad (4)$$

using the power expansion of $(1+x)^k$; the omitted terms contain terms of at least order $k^3 s^3$. Because γ_i and β_i are at most 1, sampling fractions will be substantially less than 1 at the larger venues at which many men can be sampled, and α_i will be small if the venue sampling frame is large enough, we expect the s_i to be much less than 1. If the number of sampling events is not too large, ks is small, and the remaining terms are of order $(ks)^2$ or smaller. Because the higher-order terms are small compared to ks , p is approximately equal to ks , and p is proportional to s provided that the person attends locations in the sampling frame throughout the study period.

The information necessary to obtain approximate estimates of γ and α can be obtained in the interview. The value of β for the day on which a person is sampled is known. The mean or overall sampling fraction at a venue is computed from the numbers sampled and estimated to be eligible at each sampling event. If sampling fractions at larger venues remain small near the end of the study, it seems plausible to conclude that ks is small enough that p is proportional to s . Therefore, we use $1/s$ as our weight.

DATA ANALYSIS EXAMPLE

We illustrate the effects of weighting, variation among the weights, and clustering on prevalence estimates, on estimates of the association between prevalence and a risk

factor, and on the standard errors of estimates using data on MSM from four of the six geographical areas (465–543 men in each area) in which YMS Phase II was conducted. We show that these effects may vary among areas.

Investigators asked men how often they attended bars or dance clubs during the last 6 months: never, less than once per month, once per month, two to three times per month, once per week, two to three times per week, or daily. We computed each man's value of γ (the daily probability that he went to some venue) from these data. In the following analyses, we excluded those who answered "never" because there were relatively few of them, it is difficult to assign them an appropriate value of γ , and the value would be quite small (leading to a large weight, and perhaps undue influence on a point estimate). We assigned values of γ to the remaining categories of 0.01, 0.03, 0.08, 0.14, 0.35, and 1, respectively; these values are approximately the probability of attending a bar or dance club on a day in a week (e.g., if a man attended two or three times per week, his value of γ would be 0.28 [2/7] or 0.42 [3/7], respectively; 0.35 is the average of these values). Recall that analysis results are unchanged if all weights (hence estimates of probabilities of selection) are multiplied by a constant. The interview did not contain more detailed information on either the frequencies with which a man attended each type of venue or on the frequency with which he attended the venue at which he was sampled. As a result, we must assume that all men sampled within a geographic area have the same attendance probability α the probability that a man attended the venue at which sampling took place, given that he went to some venue; we suggest in the Discussion how to avoid this assumption by collecting more detailed attendance data. Let f be the overall sampling fraction at the venue at which a man was sampled. The sampling frame data needed to compute the proportion β in Eq. 2 (the proportion of potential sampling time during which data were collected) for the event at which a man was sampled are not available at this time. Therefore we use the weight $1/(\gamma f)$ for this man.

Tables 1, 2, and 3 summarizes characteristics of the 602 sampling events at the 124 venues at which sampling occurred in these four areas. The number of venues

TABLE 1 Descriptive statistics for sampling events in the Young Men's Health Study Phase II, by metropolitan area (excluding men who reported "never" going to a venue during the previous 6 months)—number of venues, by total number enrolled and number of sampling events

	Metropolitan area				Total
	A	B	C	D	
Total number enrolled	452	527	498	484	1,961
Number of sampling events	Number of venues				
1	9	13	6	4	32
2	11	7	2	5	25
3	8	5	1	1	15
4	3	4	1	2	10
5–9	5	9	2	7	23
10–14	4	1	2	4	11
15–24	0	0	5	3	8
Total	40	39	19	26	124

TABLE 2 Descriptive statistics for sampling events in the Young Men's Health Study Phase II, by metropolitan area (excluding men who reported "never" going to a venue during the previous 6 months)—number of venues, by number enrolled at a venue

	Metropolitan area				Total
	A	B	C	D	
Number enrolled	Number of venues				
1-4	14	13	4	9	40
5-9	16	8	4	4	32
10-14	5	3	1	2	11
15-19	0	4	3	3	10
20-39	1	8	3	4	16
40-59	3	3	1	2	9
60+	1	0	3	2	6
Total	40	39	19	26	124

ranged from 19 in area C to 40 in area A (Table 1). There was only one sampling event at approximately 25% (32 of 124) of the venues, e.g., nine of 40 venues in area A (Table 1). There was great variation in the number of men enrolled at the venues and in the number enrolled at a sampling event. Four or fewer men were enrolled at 40 of the 124 sampling locations (14 of 40 locations in area A; Table 2). Overall, only one man was enrolled at 10 locations and 2 at 13 (data not shown). Among the 602 sampling events, 28% yielded only one man enrolled (40 of 144 sampling events in area A); the overwhelming majority of sampling events yielded at most four men (114 of 144, or 79%, of events in area A; Table 3). In area C, at least 60 men were enrolled at three locations (Table 2), including 99 at one and 100 at another. The large proportion of those enrolled in an area who were found at a few venues (e.g., approximately 40% of the sample

TABLE 3 Descriptive statistics for sampling events in the Young Men's Health Study Phase II, by metropolitan area (excluding men who reported "never" going to a venue during the previous 6 months)—number of venues, by number of sampling events—number (percent) of sampling events, by number enrolled at the event

	Metropolitan area				Total
	A	B	C	D	
Number enrolled	Number of sampling events (percent of events in that area)				
1	40 (28)	23 (19)	54 (33)	54 (32)	171 (28)
2	35 (24)	18 (15)	34 (21)	45 (26)	132 (22)
3	24 (17)	24 (19)	24 (15)	24 (14)	96 (16)
4	15 (10)	14 (11)	23 (14)	15 (9)	67 (11)
5-9	25 (17)	35 (28)	23 (14)	32 (19)	115 (19)
10-14	5 (3)	9 (7)	1 (1)	1 (1)	16 (3)
15-19	0 (0)	1 (1)	4 (2)	0 (0)	5 (1)
Total	144 (100)	124 (100)	163 (100)	171 (100)	602 (100)

from two venues in area C) suggests that intracluster correlation could have a substantial effect on the precision of an estimated prevalence.

Table 4 summarizes information on numbers eligible and enrolled, sampling fractions, and sampling weights. The numbers eligible include men who said that they “never” attended a bar or dance club in the last 6 months (attendance information is not available for those who did not participate). The sampling fractions at the venues were computed from numerators excluding 175 enrolled men who “never” attended a bar or dance club in the last 6 months and denominators including all men deemed eligible. The medians and the quartiles of the sampling fractions f are higher in areas B and C than in areas A and D. All four areas have similar distributions of the sampling weights $1/\gamma$ based on attendance only (ignoring the sampling fractions); the weights $1/\gamma f$ tend to be lower in areas B and C than in areas A and D (since these weights and sampling fractions are inversely proportional). Within each metropolitan area, the Spearman correlation between these two weight definitions is at least 0.95. Therefore, we expect the two sets of weights to yield similar point estimates.

Table 5 shows, for each area, the mean venue sample sizes \bar{n} , coefficients of variation (CVs) of the analysis weights, and, for each of three outcome variables, the point estimate of the kappa-type ICC, and the joint effect of the mean cluster (venue) size and the estimated ICC on the design effect. Recall that the design effect increases with $(\bar{n} - 1)ICC$ and with CV^2 . In each metropolitan area, both analysis weight CVs are approximately 1.5 (without and with including the sampling fractions in the weights). We computed estimated ICCs using Zou and Donner’s analysis of variance (ANOVA), Pearson, and kappa-like expressions⁶ which assume that, in an area, the probability of a positive result does not vary among clusters. We know that this assumption is false for HIV, as young black MSM have a higher HIV prevalence than other MSM¹ and the proportion of enrolled men who are black varies among the venues. The ICCs estimated from the Pearson and kappa-like expressions were similar and much smaller in magnitude than those computed using the ANOVA expression. We show the ICC point estimates and design effects from the kappa-like estimator. Despite the violation of the assumption given above, the ICC estimates suggest that the clustering of outcomes within venues is weak, especially for hepatitis B and unprotected anal intercourse.

Table 6 shows the point estimates, standard errors, and observed design effects for prevalences of HIV, hepatitis B, and UAI during the last 6 months if we assume that the PSU finite population correction factors are very small and can be ignored. The weighted prevalence estimates are somewhat larger than the unweighted estimates for HIV and for hepatitis B except in area B; the weighted estimates are smaller for UAI except for area C (much smaller for area A). Thus, an unweighted analysis (ignoring frequency of attendance) may produce biased estimates of prevalence in the population attending the locations in the sampling frame.

Clustering increases the standard errors of the estimates (design effects for an unweighted clustered analysis are greater than 1) for HIV and hepatitis B except for hepatitis B in area D; such an analysis of the prevalence of UAI has a design effect of approximately 1 except in area D, where the design effect is 0.4. Weighting increases the design effects, as expected, except for HIV in area C. Most of these increases are substantial, such as the increase in the design effect for the prevalence of HIV in area A from 1.2 for an unweighted

TABLE 4 Descriptive statistics for sampling at venues and sampling weights in the Young Men's Health Study Phase II, by metropolitan area (excluding men who reported "never" going to a venue during the previous 6 months)

Metropolitan area					
	A	B	C	D	All
Number eligible	874	957	897	867	3,595
Number enrolled	452	527	498	484	1,961
Sampling fractions at venues: median (quartiles)	0.48 (0.35, 0.61)	0.58 (0.50, 0.67)	0.67 (0.49, 0.89)	0.52 (0.43, 0.65)	0.54 (0.42, 0.67)
Sampling weights at venues: median (quartiles)	7.1 (2.9, 12.5)	7.1 (2.9, 12.5)	7.1 (2.9, 12.5)	7.1 (2.9, 12.5)	7.1 (2.9, 12.5)
Ignoring sampling fractions	18.0 (6.8, 37.5)	13.8 (5.4, 24.6)	14.6 (5.8, 27.9)	16.8 (5.9, 26.0)	15.3 (5.7, 27.9)
Using sampling fractions					

TABLE 5 Mean numbers of men enrolled at venues, coefficients of variation (CV) of the weights, point estimates of the intraclass correlation coefficients (ICC) for each analysis, and contribution of the estimated ICC to the design effect for each analysis, by area

Area	Mean venue size	CV of weights		HIV		Hepatitis B		Unprotected anal intercourse	
		Attendance	Sampling fraction	ICC	$(\bar{n} - 1)$ ICC	ICC	$(\bar{n} - 1)$ ICC	ICC	$(\bar{n} - 1)$ ICC
A	11.3	1.5	1.6	0.03	1.3	-0.02	0.8	0.00	1.0
B	13.5	1.5	1.5	0.08	2.0	0.04	1.5	-0.02	0.8
C	26.2	1.5	1.6	0.16	5.0	0.08	3.0	0.01	1.2
D	18.6	1.4	1.5	0.03	1.5	0.00	1.0	-0.03	0.5

Sampling fraction weights include the contribution of the attendance pattern

Mean venue size (\bar{n}) is the mean number enrolled at the venues in that area, which varies slightly for the individual analyses as a result of missing data for some enrolled men. The estimated ICC shown is from the kappa-like estimator of Fleiss and Cuzik presented in reference 6

TABLE 6 Prevalence estimates: point estimate (standard error) [design effect] in the Young Men's Health Study Phase II, by metropolitan area, by clustering, and by weighting used

Weight	Cluster	Metropolitan area			
		A	B	C	D
Proportion with HIV					
None	None	0.114 (0.015)	0.165 (0.016)	0.145 (0.016)	0.191 (0.018)
None	Venue	0.114 (0.017) [1.2]	0.165 (0.027) [2.7]	0.145 (0.048) [8.9]	0.191 (0.025) [1.9]
Attendance	Venue	0.144 (0.033) [4.8]	0.146 (0.035) [4.6]	0.184 (0.035) [4.7]	0.208 (0.040) [4.9]
Sampling fraction	Venue	0.136 (0.030) [4.0]	0.138 (0.031) [3.5]	0.190 (0.037) [5.2]	0.206 (0.038) [4.4]
Proportion with hepatitis B					
None	None	0.156 (0.017)	0.294 (0.02)	0.179 (0.018)	0.240 (0.020)
None	Venue	0.156 (0.019) [1.2]	0.294 (0.026) [1.7]	0.179 (0.035) [4.0]	0.240 (0.020) [1.0]
Attendance	Venue	0.166 (0.028) [2.6]	0.262 (0.036) [3.3]	0.219 (0.044) [6.5]	0.262 (0.043) [4.7]
Sampling fraction	Venue	0.169 (0.030) [3.0]	0.259 (0.035) [3.1]	0.238 (0.044) [6.2]	0.271 (0.044) [4.8]
Proportion with unprotected anal intercourse during the last 6 months					
None	None	0.469 (0.023)	0.417 (0.021)	0.422 (0.022)	0.526 (0.023)
None	Venue	0.469 (0.022) [0.9]	0.417 (0.021) [1.0]	0.422 (0.022) [1.0]	0.526 (0.014) [0.4]
Attendance	Venue	0.353 (0.032) [1.8]	0.408 (0.040) [3.5]	0.427 (0.042) [3.6]	0.500 (0.033) [2.1]
Sampling fraction	Venue	0.352 (0.036) [2.4]	0.404 (0.039) [3.4]	0.432 (0.040) [3.3]	0.488 (0.034) [2.2]

Weights using the venue sampling fraction also use the man's frequency of attendance

clustered analysis to 4.0 for a clustered analysis using weights incorporating sampling fractions. Using the venue sampling fractions in the weights has little effect on the prevalence estimates and their standard errors compared to the weighted results ignoring the sampling fractions. The observed design effects for the unweighted clustered analyses of the prevalences of HIV and of hepatitis B tend to be smaller than the corresponding estimates obtained from Table 5. For example, the predicted design effects $1 + (\bar{n} - 1)ICC$ obtained from Table 5 for clustered unweighted HIV prevalence estimates in areas A–D are 2.3, 3.0, 6.0, and 2.5, respectively, but the corresponding observed design effects in Table 6 are 1.2, 2.7, 8.9, and 1.9; for hepatitis B and UAI, the observed design effects in each area are smaller than those predicted from Table 5. Using PSU finite population correction factors would not affect the point estimates but would affect the standard errors and hence the design effects. For example, if we assume that each PSU finite population correction factor is 0.5, the design effect is half as large as that in the table, and the estimated standard error is multiplied by $1/\sqrt{2}$ (0.7).

The similarity between the weighted and unweighted point estimates of prevalence suggest that the proportions infected are not strongly associated with the analysis weights. We can evaluate this dependence by tabulating the proportion infected versus the attendance weights. The only “small” p values for the Cochran–Mantel–Haenszel test for trend were for hepatitis B in area B and for UAI in areas C and D (p values between 0.02 and 0.04; data not shown). If sampling probabilities are computed with less grouping, the corresponding exploratory analysis could fit a smooth curve to the binomial data. Potential smoothers are a robust loess (or lowess) curve⁷ and a kernel smoother, which can be computed in a number of statistical packages, including SAS, SPlus, stata, and R.

The largest design effect in Table 6 is for the clustered unweighted prevalence of HIV in metropolitan area C; this analysis also has the largest ICC in Table 5. We can show graphically which venues may be responsible for this clustering. Under the null hypothesis of no clustering, the observed prevalence in each venue has a binomial distribution with the proportion of persons equal to the overall (unweighted) observed prevalence. Figure 1 shows a plot of the (unweighted) observed prevalence of HIV and of UAI in each venue in area C, restricted to venues with at least 10 men enrolled, compared to exact 99% confidence intervals for prevalences of 0.145 and 0.422, respectively (the overall unweighted prevalences in these venues), given the observed sample size in the venue. For HIV as the outcome (with a design effect of 8.9 for an unweighted clustered analysis), three venues have prevalences outside these confidence intervals; one venue with approximately 40 men enrolled has a prevalence far above the upper limit of the confidence interval. In contrast, for UAI as the outcome (with a design effect of 1.0 for an unweighted clustered analysis), all observed prevalences are within the confidence intervals.

UAI is a risk factor for HIV and hepatitis B infection. Table 7 shows the effects of clustering and weighting on estimates of the associations between UAI and these infections in areas A and B from logistic regression models fit using both the SAS Proc SurveyLogistic and from generalized estimating equation models fit using SAS Proc Genmod. The two models give very similar estimates of the log odds ratio (OR) and its standard error. The design effects for clustered unweighted analyses are close to 1 but are clearly greater than 1 with weighting. Weighting has a large effect on the point estimates of the log OR for HIV, with substantial increases in area A and decreases in area B. The increase in the standard error for the log OR for the

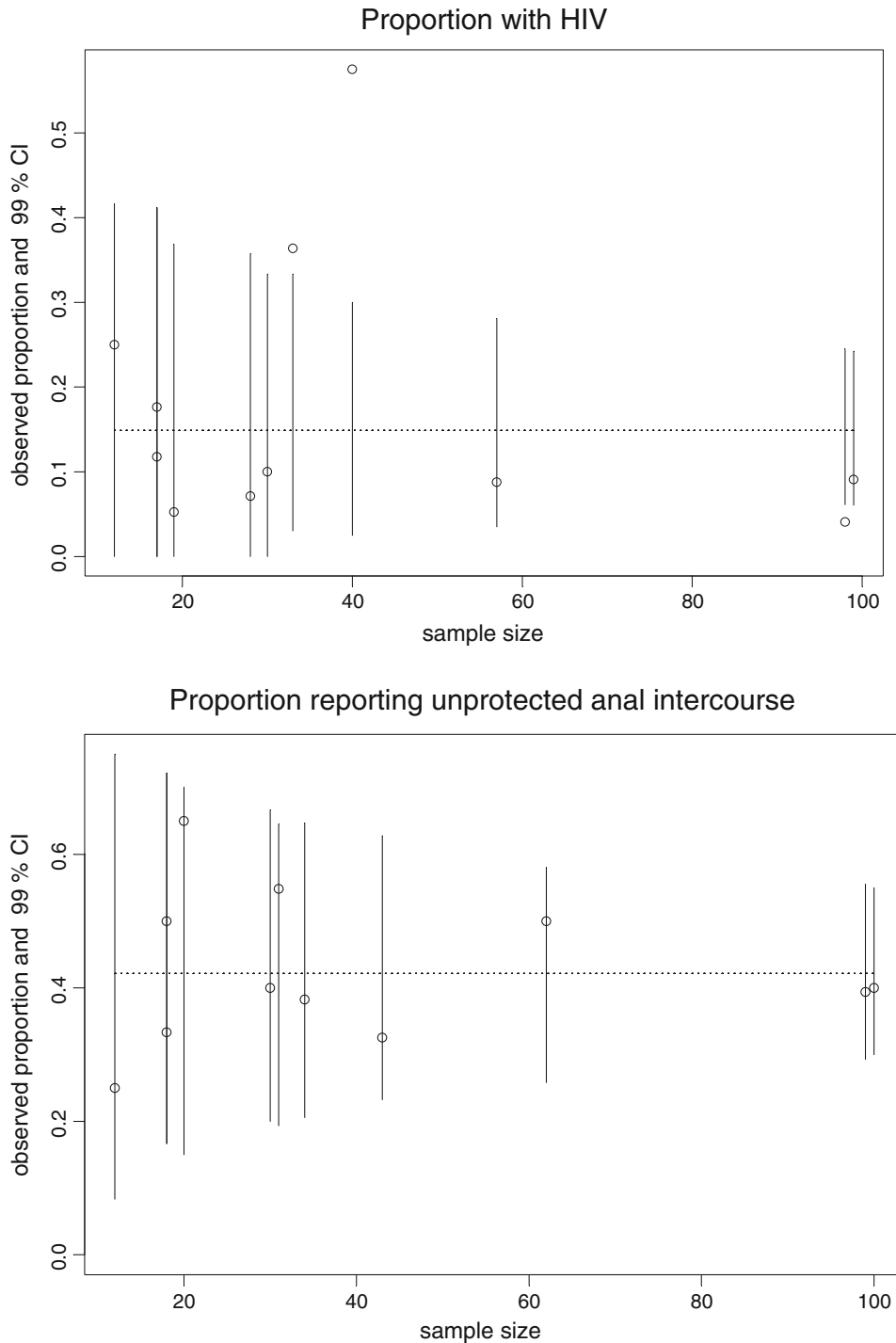


FIGURE 1. Proportions with HIV and unprotected anal intercourse during the last 6 months in area C, by venue. The *dotted line* shows the proportion for all persons combined (using unweighted analyses). For each venue, the *open circle* shows the observed proportion; the *vertical bar* shows a 99% exact confidence interval for this proportion in that venue based on the number sampled at that venue, if the true proportion is equal to the proportion for all venues combined.

TABLE 7 Results from logistic regression models for the associations between unprotected anal intercourse in the last 6 months and HIV and hepatitis B in metropolitan areas A and B: Young Men's Health Study Phase II

Area	Weight	Clustering	Survey logistic model		GEE model	
			Log OR (se) [DE]	<i>p</i> value	Log OR (se)	<i>p</i> value
Outcome: HIV						
A	None	None	0.549 (0.302)	0.07		
A	None	Venue	0.549 (0.336) [1.2]	0.10	0.544 (0.334)	0.10
A	Attendance	Venue	0.752 (0.597) [3.9]	0.21	0.743 (0.581)	0.20
A	Sampling fraction	Venue	0.917 (0.620) [4.2]	0.14	0.909 (0.611)	0.14
B	None	None	0.177 (0.238)	0.46		
B	None	Venue	0.177 (0.282) [1.4]	0.53	0.109 (0.289)	0.71
B	Attendance	Venue	0.092 (0.435) [3.4]	0.83	0.026 (0.421)	0.95
B	Sampling fraction	Venue	0.013 (0.429) [3.3]	0.98	−0.06 (0.427)	0.89
Outcome: hepatitis B						
A	None	None	0.168 (0.264)	0.52		
A	None	Venue	0.168 (0.243) [0.8]	0.49	0.162 (0.237)	0.50
A	Attendance	Venue	0.338 (0.519) [3.9]	0.51	0.336 (0.515)	0.51
A	Sampling fraction	Venue	0.400 (0.590) [5.0]	0.50	0.398 (0.587)	0.50
B	None	None	0.495 (0.193)	0.01		
B	None	Venue	0.495 (0.162) [0.7]	0.00	0.484 (0.160)	0.00
B	Attendance	Venue	0.399 (0.294) [2.3]	0.17	0.376 (0.294)	0.20
B	Sampling fraction	Venue	0.449 (0.289) [2.2]	0.12	0.430 (0.291)	0.14

GEE generalized estimating equation, SE standard error

Weights using venue sampling fraction also use the man's attendance pattern

weighted hepatitis B analysis in metropolitan area B causes a dramatic increase in the *p* value for the association with UAI, from a highly significant association to a clearly nonsignificant association. Examples of code to compute the results in Tables 6 and 7 are in the [Appendix](#).

DISCUSSION

Special methods are required for sampling hard-to-reach populations (those for which there is neither a list of the population members nor a list of where they can be found). Kalton provided a recent review of methods for sampling these populations.⁸ The most popular methods for sampling populations such as MSM and injection drug users are time–location, respondent-driven, and snowball sampling. Two recent reviews contain substantial discussions of time–location and respondent-driven sampling.^{9,10} We provide a brief comparison of these two methods at the end of the “[Discussion](#)” section.

Reports of TLS data analyses should clearly state the population to which the estimates apply. If the data are analyzed as a simple random sample, the results apply to a population resembling the persons actually sampled, corresponding to data from a surveillance system. If only the sampling fractions *f* in Eq. 3 are used to compute weights for prevalence (frequency of attendance at venues and the proportion of potential time used in sampling [γ , α , and β in Eq. 3] are not used), or if the analysis is unweighted but the sampling fractions are approximately constant, then prevalence point estimates refer to the population of visits to the locations in the sampling frame, ignoring the fact that there is variation in the

frequency with which persons appear at those locations. Persons who attend locations in the sampling frame more frequently are over-represented in the sample compared to the population of all persons who attend these locations. Therefore, if there is a positive association between frequency of attendance and an outcome of interest, the prevalence estimate for this outcome is likely to be larger than the prevalence in the population of persons attending these locations. Unless the possibility of clustering within locations is accounted for, the confidence interval for a prevalence estimate from such an analysis may be too narrow (see Table 6 and the discussion below), provided that we can ignore the finite population correction factor for the sampling of locations (the primary sampling units).

The CDC recently published two reports on HIV in MSM, based on data from the National HIV Behavioral Surveillance (NHBS) system.^{11,12} The initial publication¹¹ gives prevalence estimates in individual cities; the second publication¹² uses logistic regression to examine associations between risk factors and HIV infection, aggregating over geographical areas. The NHBS system uses time-location sampling to collect data.³ The published results regard the data as simple random samples (clustering within venues and cities, differential venue attendance patterns, and sampling fractions at individual sampling events are not considered; the logistic regression models do include geographic area as a fixed effect). As summarized above, estimates from these analyses describe a population resembling the actual men sampled, rather than a larger population. Based on the results in Tables 6 and 7, we expect that the published confidence intervals would be narrower than confidence intervals for estimates pertaining to the population of MSM attending venues in the sampling frame. CDC staff fit a variety of models that account for clustering at the venue level, including a generalized linear mixed model and a generalized estimating equation model. For these models, the lengths of the confidence intervals for the log odds ratios differed from those from the published model by at most 2% and approximately 10%, respectively. Therefore, for these data, conclusions from logistic regression models seem not to be affected by clustering. However, such an analysis would be essential before ignoring the possible effects of clustering in the analysis of TLS data from other studies. CDC is currently reviewing TLS weighting strategies, similar to those presented here but specifically tailored to the NHBS protocol, for possible use in NHBS data analyses.

In contrast, an analysis using weights based on the probability that a person is sampled, accounting for the variation in the frequency with which persons attend a location in the sampling frame, yields estimates that refer to the population of persons who attend these locations. Precision estimates are invalid unless they allow for the effect of clustering within sampling locations on standard errors. Accounting for clustering does not affect point estimates.

Our results show that using weights and accounting for clustering can yield both point estimates and standard error estimates that are substantially different from those obtained from regarding the data as arising from a simple random sample (Tables 6 and 7). If we could collect better data on venue attendance (as suggested below), we could estimate α (the probability of attending the location where sampling took place) in Eq. 3; we suggest an estimate for γ (the probability of attending some location in the sampling frame) which would increase the variation of γ among persons. If some location sampling periods did not include all potential sampling times at a location, some of the terms β_i in Eq. 3 would not be 1. We would use the inverse of $\gamma\alpha\beta_i$ as the participant's analysis weight. Increased variability in the first three terms in this product would increase the variability in the

weights. Because the design effect increases with the coefficient of variation in the weights, more complete data would likely result in larger design effects (larger standard errors) for clustered weighted analyses than those in Tables 6 and 7.

Our proposed weights are proportional to the inverses of estimated sampling probabilities for the study. These estimated probabilities rely on assumptions that would be hard to verify and on self-reported data, resulting in weights that can only be approximations to the correct weights. To evaluate the potential for biased estimates as a result of incorrect weights, the data analysis should include an evaluation of associations between weights and both outcomes of interest and important covariates, as we suggested in our analyses of the YMS Phase II data. If there are clear associations, appropriate care should be used in drawing conclusions from the study. If some persons are assigned relatively large weights, perhaps as a result of being one of only a few persons sampled at a location, it may be desirable to check the robustness of the results by redoing the analysis after those persons are deleted from the data. Alternatively, one may consider weight-trimming methods that reduce the influence of persons with extreme weights. Because our weights are approximations to the correct weights, the true standard errors of our estimates are likely to be larger than the computed standard errors. As a result of the possible bias in a point estimate and underestimation of its standard error, conclusions from analyses of TLS data should be made cautiously.

We ignored both the fpc and chose to combine all sampling events at each location in our analysis. It may be possible to compute the fpc if the analysis is based on individual sampling events, rather than combining all events at each location. In YMS Phase II, at most three men were enrolled during at least half of the sampling events in each of the four metropolitan areas (Table 3). An analysis using individual sampling events would produce unstable estimates as a result of these small numbers. In addition, it would ignore the correlation between men sampled at different times at a location, possibly leading to an underestimate of variance (consider the case when all men at a location have the same response, so the sample size contributed to the analysis by that location should effectively be 1, not the number of men sampled or the number of times the location was sampled). Because the validity of the weights is uncertain, resulting in potential bias and underestimation of standard errors, we prefer to have conservative confidence intervals for our estimates. Ignoring the fpc provides such intervals. Computing the fpc with multiple sampling events at a venue is an unsolved problem.

If reliable data can be obtained on the frequency with which a person is present at each type of location in the location sampling frame, we can make better estimates of a person's sampling probability. Suppose that such data can be collected. Analogous to Eq. 3, let γ be the probability that the person is present at some location of the type at which he was sampled. For each location i of this type, let n_{ij} be the number of eligible persons counted at the j th sampling event at this location, and let \bar{n}_i be the average of n_{ij}/β_{ij} (β_{ij} is defined as in Eq. 3) over all sampling events at this location (\bar{n}_i is the estimated average attendance at location i during periods when sampling could have occurred). If the location at which the person was sampled is indexed by $i=1$, let $\alpha = \bar{n}_1/\sum \bar{n}_i$ (this is an estimate of the probability that the person went to the location where he was sampled, given that he went to a location of this type). Let f be the sampling fraction at this venue based on all sampling events combined (we prefer to combine sampling events, as the numbers of eligible and sampled persons at individual sampling events may be small). Let β be the proportion of possible sampling hours on this day at this location during which

data were collected. We propose using the inverse of $\gamma\alpha\beta f$ as the analysis weight. We need appropriate studies of the feasibility of collecting detailed data on the frequency with which men attend venues of different types and of the accuracy of such data.

If we wish to make inference to the population from which persons attend sampling locations, we emphasize that it is essential to collect data that allow us to compute an approximation to the probability that a person is sampled. We outlined above the location attendance information that would permit estimation of γ and α . We also need estimates or counts of the numbers of eligible persons. If nonresponse may depend on a characteristic that can be obtained from eligible nonrespondents that is also associated with an outcome or covariate of interest, this characteristic should also be recorded for all eligible persons who are counted. The details of the sampling frame used to schedule sampling events must be available in order to compute β . The length of time during which eligible persons were counted should also be recorded, as β should be computed by dividing this duration by the total time potentially available for sampling at that location on that day. The sampling frame details would also be needed if theory is developed to use the finite population correction factors for the primary sampling units in estimating standard errors.

As noted earlier, analysis weights should incorporate the probability that we obtain data from a sampled person (in our case, a man asked to participate) by multiplying the estimated selection probability by the estimated response probability. In our example, if men with a particular characteristic were more likely to cooperate, they were over-represented among respondents, possibly biasing our estimates. It is common practice in creating survey analysis weights to incorporate not only the design weight, which reflects an individual's chance of selection, but also a response probability (to adjust for nonresponse). This procedure assumes that the factors associated with the probability of response (typically demographic characteristics such as age, race/ethnicity, and sex) are related to the outcomes of interest. In a population survey, the response probability may be able to be computed from the known demographic characteristics of the population. To implement this procedure in a TLS survey such as YMS II, interviewers would need this information for those who refuse to participate; because many or most may refuse to provide information, it may be feasible to use only characteristics that can be estimated from appearance, such as age range or race. We incorporated response probabilities in our definition of the sampling fraction f . If response probabilities depend on one or more demographic characteristics, there are two possibilities. The sampling fractions could be computed by stratifying on these characteristics. Alternatively, it could be assumed that sampling fractions do not depend on these characteristics, and a separate multiplicative proportion could be included in the expression for the weight. In the latter case, we propose combining sampling events at a location to compute response probabilities for the same reasons as for computing location sampling fractions.

Our analysis recommendations have assumed that sampling fractions and response probabilities at a location do not depend on time, and that the data provide good estimates of these proportions even at locations with only a few persons sampled. It may be desirable to consider grouping some locations at which few persons are sampled with other locations (in YMS II, locations of the same type) to estimate these proportions; logistic regression could be used to evaluate this possibility. It may also be desirable to check whether these proportions vary over time by plotting proportions against time and by a categorical data analysis.

Design effects may be large and probably cannot be estimated in advance (Tables 6 and 7). Variation in the weights increases the design effects. Sample sizes

for a time–location sample must take into account possible large design effects. Our analyses highlight the need to account for design effects greater than 1 as a result of clustering and variation in the weights, in addition to potential sampling bias, when analyzing TLS data. The importance of design effects greater than 1 in respondent-driven sampling (RDS) analyses has been discussed,¹³ but to our knowledge has rarely been considered in the analysis of TLS data.

Finally, the primary alternative to TLS is RDS.¹⁴ Two studies in Brazil found substantial differences between the estimates from these two sampling methods.^{15,16} One study of MSM shows that these two methods may yield samples with very different properties with respect to a characteristic of interest (in this case, socio-economic status) that may be associated with an outcome of interest.¹⁵ Inference from TLS is limited to the population of persons attending venues in the sampling frame where sampling took place; in a report on TLS study, it would be desirable to summarize characteristics of locations at which no persons could be recruited to assess possible differences between the reachable population of venue attendees and the complete population of venue attendees. TLS is also likely to be slower and more expensive than RDS.¹⁵ We have outlined the assumptions required to compute valid analysis weights for a sample survey analysis of TLS data. Theory for RDS depends on assumptions concerning respondents' knowledge of their network, the way respondents recruit others, the character of the network, and the recruitment process;¹⁷ see also Gile and Handcock's recent critical evaluation.¹⁸ Custom software is available to implement RDS (www.respondentdrivensampling.org). See Semaan¹⁰ for further discussion of software and evaluation. Using an RDS sample of a population with known parameters, Wejnert¹⁹ showed that current methods for estimating variance are not consistent and may over or under estimate uncertainty. Similarly, using simulation, Goel and Salganik¹³ showed that, under certain conditions, current RDS software reports confidence intervals that are substantially too narrow. In addition, the current RDS theory does not cover fitting linear and logistic regression models.

The choice between TLS and RDS should consider ease of implementation, time needed to complete the study, cost of the study, coverage of the population of interest, assumptions required for the analysis to be valid, data required for the analysis, difficulty in doing the analysis, and feasibility in doing the analysis. Because our proposed TLS analysis is based on standard sample survey methods, compared to RDS it has the advantages of a firmer theoretical basis, the ability to fit models, and the use of widely available software. Both methods have design effects greater than one; further research is needed to determine which, if either, method has a consistently smaller design effect than the other.

ACKNOWLEDGMENTS

We thank Christopher H. Johnson, Nevin Krishna, Lillian S. Lin, Alexandra Oster, and Ryan E. Wiegand, Centers for Disease Control and Prevention (CDC), for helpful comments on the manuscript. John Karon's work was done as a contractor for CDC.

APPENDIX: COMPUTATIONAL SOFTWARE

We analyze a time–location survey as a two-stage cluster sample, sampling of venues (PSUs) and sampling within venues (the second stage). The number of persons

sampled varies among venues, and persons are sampled with unequal probabilities. The sample should be analyzed using software that can accommodate such a sampling design.

Many software packages could be used, some of which are free. The website www.hcp.med.harvard.edu/statistics/survey-soft/ contains brief descriptions of and links to many of these packages. Well-known commercial packages that could be used include SUDAAN, SAS, Stata, and SPSS (with an additional module). We show code from SAS, which produced the results in Tables 6 and 7 and from R, a free package.

We assume that the dataset has one row for each person sampled. As an illustration, we assume that for each person we have the following variables: Site, the area in which sampling was done, if there were multiple areas, as in our data; Stratum, a code for stratum within site—the probability of sampling a PSU is uniform within each stratum; Venue, the venue at which sampling was done; Weight, the sampling weight, the inverse of a constant times an estimate of the probability that the person was sampled; Status, an indicator variable for the presence of the condition for which prevalence is to be estimated (1 if the condition is present); and RiskFactor(s), one or more variables with data on a risk factor for which the association with the condition is to be estimated (in Table 7, unprotected anal intercourse in the past 6 months). In addition, if the sampling probability for PSUs was not constant, for analyses in SAS we need a second data set with one row for each site and stratum combination, with variables Site, Stratum, and Rate, where Rate is the proportion of PSUs sampled in that stratum. Status, Weight, RiskFactor(s), and Rate should be numeric.

Analysis in SAS

We assume these data sets are SAS data sets with names DataPersons, and, if sampling probabilities vary among PSUs, DataRates, in the folder c:/SAS/data. To obtain the estimated prevalences and standard errors in Table 6, we used the following code:

```
LIBNAME in "c:/SAS/data";
Data Step1; set in.DataPersons; run;
Proc sort data=Step1; by Site; run;
Proc SurveyMeans data=Step1 Rate=<value or DataRates>mean stderr var clm;
by Site;
cluster Venue; Var Status; weight Weight;
ODS output Statistics=SummaryData;
run;
```

If all PSUs are sampled with the same probability, Rate should equal the sampling probability, and the cluster statement should be omitted. The ODS statement uses the Output Delivery System to create the data set SummaryData with one row for each value in site. For each site, the dataset contains the estimated prevalence (mean), and its standard error (stderr), variance (var, useful for computing design effects), and a confidence interval (clm, by default, a 95% interval). The SurveyMeans procedure produces several pages of output for each site.

The code for logistic regression results uses the SurveyLogistic procedure available in SAS version 9. It is easiest to do the analysis for a single site. Let step2 be the data set with data for one of the sites. If sampling probabilities vary among

PSUs for this site, let `DataRatesSite` be the data set restricted to this site.

```
Proc SurveyLogistic data=step2 Rate=<value or DataRatesSite>;
strata Stratum; cluster Venue; weight Weight;
ODS output ParameterEstimates=ParameterEsts;
model Status (desc)=RiskFactor;
run;
```

The ODS statement creates a data set with parameter estimates. If there are repeated analyses, the data sets can be concatenated to summarize the output.

Analysis in R

R is a powerful statistical programming package that implements the S language²⁰ and contains functions to implement many statistical procedures. The package can be downloaded from <http://cran.r-project.org/>; you will also need the library named `survey`, obtained from a link on the same home page. The Help feature on the R commands page has links to two useful documents, *An Introduction to R*, and *R Data Import/Export under Manuals*. See Venables et al.²¹ for an introduction to R; a preliminary 114 page version is available from the R website.²² Another useful document is *A Survey Analysis Example*,²³ by Thomas Lumley, author of the R `survey` library; see also Lumley's book.²⁴ The following discussion assumes that a user has learned the basics of R.

Survey Data Analysis The structure of the dataset is the same as for SAS, except that there should be an additional variable `FPC`, which is either the number of venues in the sampling frame from which the venue was chosen, or the sampling probability for the venue. If the venue was chosen with certainty, `FPC` should be 1. To ignore the finite population correction, set `FPC` to a small number (such as 0.001). Persons with a missing value for essential data (sampling weight, venue) should be removed. If the dataset is ASCII, it should be edited to change the missing value code to `NA`; otherwise, use a distinctive numeric code for missing data. It is desirable, but not necessary, to have the variable names in the first row of the data set. It is easiest if this dataset is saved in the folder containing the R software.

We first define a data frame containing the data. If the data set is in a project folder, replace "filename" by the path and filename. The operator "<->" defines an object to be the value of the function or expression on the right-hand side of this operator.

```
library(survey) # load the survey library
DataFrame<- read.table("filename", header=TRUE) # variable names in data set
DataFrame<- read.table("filename", header=FALSE, col.names=c(character list)
# variable names not in row 1 of data set
# character list is "variable1", "variable2", ..., "variablev"
DataFrame[1:10,] # look at the data for the first 10 rows of the data set
```

We now analyze the prevalence of `Status`. We use one of the first three assignments below to create a new data frame from which persons with `NA` are removed. The operator "!" is the logical operator "not". Use the third assignment if the missing value has the numeric value `Mcode`.

```
DataStatus<- DataFrame[!is.na(DataFrame[, "Status"]),]
DataStatus<- DataFrame[(DataFrame[, "Status"] !=NA,]
```

```
DataStatus<- Dataframe[(Dataframe[, "Status"] !=Mcode,]
# create an object with the survey design information (order of function #
arguments is arbitrary)
SurveyStatus<- svydesign(data=DataStatus, ids=~Venue, strata=NULL,
weights=~Weight, fpc=~NFrame)
svymean(x=~Status, design=SurveyStatus, deff=TRUE) # prevalence estimate
```

In the previous two statements, note that it is *essential* to have the tilde (“~”) in defining some function arguments. `svymean()` returns the variable analyzed (Status), the weighted mean, the standard error, and the design effect. Omit the specification of the design effect if it is not to be computed.

To obtain the logistic regression estimate of the association between Status and RiskFactor, we first remove persons with missing values of RiskFactor, e.g.

```
DataStatusRisk<- DataStatus[!is.na(DataStatus[, "RiskFactor"]),]
SurveyStatusRisk<- svydesign(data=DataStatusRisk, id=~Venue, weights=~Weight,
fpc=~NFrame)
logitmodelStatus<- svyglm(Status~RiskFactor, design=SurveyStatusRisk,
family=quasibinomial())
summary(logitmodelStatus) # results of logistic regression
```

`summary()` prints information about the model used and, for the intercept and each risk factor, the regression coefficient, standard error, Student’s *t* value, and *p* value.

REFERENCES

1. Valleroy L, MacKellar DA, Karon JM, et al. HIV prevalence and associated risks in young men who have sex with men. *J Amer Med Assoc.* 2000; 284(2): 198–204.
2. MacKellar D, Valleroy L, Karon J, Lemp G, Janssen R. The Young Men’s Survey: methods for estimating HIV seroprevalence and risk factors among young men who have sex with men. *Public Health Rep.* 1996; 111(Supplement): 138–144.
3. MacKellar DA, Gallagher KM, Finlayson T, Sanchez T, Lansky A, Sullivan PS. Surveillance of HIV risk and prevention behaviors of men who have sex with men—a national application of venue-based, time-space sampling. *Public Health Rep.* 2007; 122 (suppl 1): 39–47.
4. Weinbaum CM, Lyerla R, MacKellar DA, et al. The Young Men’s Survey Phase II: hepatitis B immunization and infection among young men who have sex with men. *Amer J Public Health.* 2008; 98(5): 839–845.
5. Kish L. *Survey Sampling.* New York, NY: Wiley; 1965.
6. Zou G, Donner A. Confidence interval estimation of the intraclass correlation coefficient for binary outcome data. *Biometrics.* 2004; 60(3): 807–811.
7. Cleveland WS. *Visualizing Data.* Summit, NJ: Hobart; 1993.
8. Kalton G. Methods for oversampling rare subpopulations in social surveys. *Surv Methodol.* 2009; 35(2): 125–141.
9. Marpsat M, Razafindratsima N. Survey methods for hard-to-reach populations: introduction to the special issue. *Methodol Innov Online.* 2010; 5(2): 3–16.
10. Semaan S. Time-space sampling and respondent-driven sampling with hard-to-reach populations. *Methodol Innov Online.* 2010; 5(2): 60–75.
11. Centers for Disease Control and Prevention. Prevalence and awareness of HIV infection among men who have sex with men—21 cities, United States, 2008. *Morb Mortal Wkly Rep.* 2009; 59(37): 1201–1207.

12. Oster AM, Wiegand RE, Sionean C, et al. Understanding disparities in HIV infection between black and white MSM in the United States. *Epidemiol Soc.* 2011; 25(8): 1103–1112.
13. Geol S, Salganik MJ. Assessing respondent-driven sampling. *Proc Natl Acad Sci.* 2010; 107(15): 6743–6747.
14. Heckathorn DD. Respondent-driven sampling: a new approach to the study of hidden populations. *Soc Probl.* 1997; 44(2): 174–199.
15. Kendall C, Kerr LRFS, Gondim RC, et al. An empirical comparison of respondent-driven sampling, time location sampling, and snowball sampling for behavioral surveillance in men who have sex with men, Fortaleza, Brazil. *AIDS Behav.* 2008; 12(suppl 1): 97–104.
16. McKenzie DJ, Mistiaen J. Surveying migrant households: a comparison of census-based, snowball and intercept point surveys. *J R Stat Soc A Stat Soc.* 2009; 172(2): 339–360.
17. Volz E, Heckathorn DD. Probability based estimation theory for respondent driven sampling. *J of Official Stat.* 2008; 24(1): 79–97.
18. Gile KJ, Handcock MS. Respondent-driven sampling: an assessment of current methodology. *Soc Methodol.* 2010; 40(1): 285–327.
19. Wejnert C. An empirical test of respondent-driven sampling: point estimates, variance, degree measures, and out-of-equilibrium data. *Soc Methodol.* 2009; 39(1): 73–116.
20. Becker RA, Chambers JM, Wilks AR. *The New S Language: A Programming Environment for Data Analysis and Graphics.* Pacific Grove, CA: Wadsworth & Brooks/Cole;1988.
21. Venables WN, Smith DM, and the R Development Core Team. *An Introduction to R, second edition.* (No city given) United Kingdom: Network Theory Limited,; 2009.
22. <http://www.cran.r-project.org/doc/contrib/Verzani-SimpleR.pdf>. Accessed 21 April, 2011.
23. <http://faculty.washington.edu/tlumley/survey/doc/survey.pdf>. Accessed 21 April, 2011.
24. Lumley T. *Complex Surveys: A Guide to Analysis Using R.* Hoboken, NJ: John Wiley & Sons, Inc.; 2010.