November 1998

A Probability Sample of Gay Urban Males: The Use of Two-Phase Adaptive Sampling

Address correspondence to:

Johnny Blair, Survey Research Center, 1103 Art/Sociology Bldg., University of Maryland, College Park, MD. 20742;
e-mail: johnny@cati.umd.edu

The advantages of probability sample designs over convenience samples have been known since the 1930s (Neyman, 1934). Strictly speaking, in a probability sample every member of the target population has a known, non-zero chance of selection. This provides a statistical basis for projecting sample estimates back to the target population.

In studies of relatively rare populations which are also not easily identifiable, such as gay men, there are serious obstacles to the implementation of probability designs. The population's rarity means that large numbers of households must be contacted to locate the target sample, which greatly affects the survey’s cost. The reluctance of many gay men to report their homosexuality in a survey -- particularly at the very start of an interview -- can cause severe undercoverage and potential bias. That is, many members of the population are never identified and those that are may differ substantially in their characteristics from the population at large.

These obstacles have often led to the use of extremely loose survey methods, such as sampling patrons of gay bars or members of gay rights organizations. These convenience samples provide no statistical basis for projecting back to the population of interest.

The major goal of the present study was to provide reliable (i.e., replicable) population estimates for the gay male population

of four major cities. In order to implement the probability sample necessary to achieve this objective, a survey sample design was selected that took into account the likelihood of flaws in the data used for planning.

Typically, in a sample survey there is sufficient initial information available about the target population size and the sampling frame to specify completely such sample design parameters as stratification definitions, within-stratum sample sizes, sampling rates, and the geographic distribution of the sample. However, when the initial information is incomplete or may be unreliable, it can be useful to make these specifications only provisionally, with their final form dependent on information obtained during sample selection and data collection. When, in addition to these uncertainties, the target population is a small fraction of the general population, the risks to the study costs and the sample size ultimately achieved may be high. That is, errors in the initial assumptions may make the survey much more costly, resulting in fewer interviews being possible for a given fixed budget.

Overestimation of the population prevalence, based on secondary sources, occurs for two reasons. First, such sources usually only approximate the definition (or are indicators of the presence) of the target population, for example, the number of single males in a particular age range or the number of reported

AIDS cases. Second, even when a source is a direct estimate of the target group, the identification often is not from a survey. When the sources are not based on survey data, they do not account for underreporting, such as the fact that some members of the target population will deny their eligibility.

Multiple secondary sources are often used and these sources differ in their accuracy. Moreover, data from these sources must be combined into a single estimate of prevalence. This estimation requires some judgment by the researcher, who is, in effect, modeling prevalence. Because of these problems, important areas of sample design focus on efficient methods for locating rare (or moderately rare) or elusive populations. Sudman, Sirken, and Cowan (1988) provide an excellent overview of many of these designs. Sudman (1985) gives a detailed analysis of the balance between costs and variances. One class of such sample designs takes advantage of the natural geographic clustering of the target population.

In this paper, the key aspects of a two-phase, adaptive sampling approach are described for a telephone survey of self-identified gay urban males[1]. As noted, the study objective required obtaining a probability sample of gay males in the central cities of four urban areas: San Francisco, Chicago, New York, and Los Angeles. Households were screened and data was collected by telephone interview with identified eligible respondents. Simply

screening large numbers of households in search of the target group was clearly prohibitively expensive. The overall expected eligibility -- hit rates -- in these areas, while not known exactly, were expected to be in the range of 3% (Binson et al, 1996). To obtain the desired 4,600 interviews, over 150,000 households would have had to be screened. When contact and cooperation rates are factored in, that number would increase considerably.

In sampling even moderately rare populations, errors in estimated hit rates that would be inconsequential with more prevalent populations can be disastrous. If a target group makes up 50% of the general population, an error in that estimate of -3% means that the necessary sample size increases by about 6%. When the target group is expected to constitute only 5% of the general population, a 3% negative error more than doubles the required sample size. Of course, a positive error has an effect of equal magnitude in the opposite direction. But the risk is in underestimation of the hit rate.

It was known from past surveys and other data that this target population, to some extent, clusters geographically. For example, there are known gay neighborhoods; however, it is not advisable to limit sampling to identified neighborhoods, since large numbers of the target group may not be covered. Additionally, to the extent that gays outside known neighborhoods differ on study variables,

seriously biased estimates could result. Based on discussions with local informants, it is suspected that some amount of residential clustering of gay males also exists outside known neighborhoods. In such locations, the hit rates, while not as high as in the known neighborhoods, should be much higher than for the urban areas as a whole. To take advantage of the naturally occurring geographical clustering requires knowing the location of these areas, defining their boundaries clearly, and estimating the likely hit rates. The data available to design such studies may be insufficient or of questionable accuracy.

METHODOLOGY

Blair and Czaja (1982) showed how a population that is highly clustered may be efficiently sampled in two phases for a telephone survey. In their study of black households, a first-phase simple random sample of telephone numbers was selected and called to determine eligibility[2]. The expectation was that black households would be identified at a rate approximately equal to their occurrence in the population. This was confirmed. Differences were expected due to some households refusing to complete the screening interview, nonrandom errors of inclusion or exclusion, and sampling variance. Since phone numbers are to some extent assigned on a geographical basis, even within area codes, numbers numerically close to each other often represent

households which are geographically close together. In the Blair and Czaja study (1982), each first-phase phone number connected to an eligible household was used to define a bank of 100 telephone numbers. For example, if 312-996-5693 was identified as a black household, then the 100 numbers from 321-996-5600 to 321-966-5699 were defined as a bank. Calling then continued in each bank until a prespecified number of eligible households was identified. It was found that within these second-stage banks eligible households occurred at twice the rate of their occurrence in the general population.

This two-stage cluster design produces an equal probability sample of eligible households. In the first stage, banks are selected with probabilities proportional to the number of black households in the bank. The probability of selection of a bank is:

$$P_B = \frac{N_i}{\sum N_i / m}$$

where N_i is the number of eligible households in bank i, $\sum N$, is the total number of black households in all banks, and m is the number of banks to be selected.

In the second stage, numbers are called within each bank until a fixed number, k, of additional eligible households is identified. The cluster size k is the same for all second-stage banks. The second stage probability is:

$$P_{ri} = \frac{k_i}{N_i}$$

Taking the two stages together produces individual inclusion probability

$$P_R = \sum \frac{N_i}{N_{i/m}} \times \frac{k_i}{N_i} = \frac{n}{N}$$

where n = mk = $\sum k_i$. That is, all target households have the same chance of selection, regardless of the population's prevalence in any given bank.

This design was used effectively in other studies (see Inglis, Groves, and Heeringa, 1987) of populations that cluster geographically. Waksberg (1983) pointed out a potential weakness in the design that arises when it is not possible to reach the cluster size k in many clusters. In that situation, completely unbiased estimation requires that short clusters be weighted. The effect of the weight is to increase the sampling variance. In some cases, the variance might be increased to such an extent that the design, taking into account both cost and variance, is no more efficient than a straight screening design. As will be seen, an important design factor is to determine in which circumstances the cluster size will be achievable, and to examine alternative estimation (i.e., weighting) strategies when it is not.

Two things can be done to address this concern. First, k

must be chosen judiciously, so that while it is large enough to take advantage of the clustering, not too many short clusters result. Second, extreme weights can be trimmed to reduce their effect on sampling variances, while still producing nearly unbiased estimates. Additionally, it is worth noting that not all sample surveys are primarily concerned with producing estimates of population parameters. In many AIDS surveys, the main analysis concerns the relationships between variables, such as the relationship between the respondents' knowledge and their behaviors. Often in these types of analyses weights are not used. However, a sufficient number of cases is necessary for the analysis. A sample design that efficiently increases the yield of target cases can be valuable for such analyses, even if it is only marginally useful for estimating population parameters. While these sorts of designs have been shown to work with highly clustered groups such as black households, it is not clear how effective this approach is when clustering is thin, such as in the present study.

The sample design used for the present study was a two-phase design with adaptations based on data collection experience. The basic notion of adaptive sampling (Thompson, 1992) is a simple one: to use data from early observations in a survey to guide selection of additional elements. Adaptive sampling in this study used data collected on household eligibility to determine

whether to (a) redefine stratum boundaries, (b) determine within-stratum sampling procedures, (c) adjust cluster sizes, (d) drop or undersample costly strata, or (e) reduce or redistribute target sample sizes.

The method of screening both for eligible households and for data collection, was the telephone. Once an eligible household was identified, a respondent was randomly chosen from among the household's eligible residents. In the first phase, a list-assisted Random Digit Dial (RDD) sample was selected in each stratum. This unclustered design gives each residential telephone number an equal chance of selection. In phase two, it was expected that for some strata the hit rate would be so low that no substantial gains would result from clustering. It is actually disadvantageous to cluster in such situations. However, in strata where the first-phase hit rate was reasonably high, there seemed more likelihood that there might be some within-stratum clustering that could be exploited by using a Mitofsky-Waksberg sample. The danger of falling short of the cluster size is, of course, still an issue even in these select strata. In addition to the strategies described above, a protection afforded here is that because of the low target population density, the cluster size is very small. The increase in sampling variances due to weighting is exacerbated when there is large variation in the size of the weights; for example, some clusters miss the

target k by a lot, while others exceed it[3]. When the cluster size is very small, the range of sizes of possible weights is automatically constrained. Still, whether the anticipated yield would approximate the number needed for cost-effective screening was unknown, so the adaptive sampling options remained centrally important.

Working with imperfections in both the sampling frame and screening data, the design sought to balance several factors: yield (target sample size), cost, and variance. The resulting design was based on cluster sampling in conjunction with adjustments made during the data collection precess, while still maintaining a probability sample. An extensive effort was made to identify sampling locations in each of the four cities and to collect and collate data from a very wide variety of sources about the expected occurrence of the target group within each location (Binson et al., 1996). These sources included census data, AIDS cases reported by health departments, marketing lists, and directories of gay businesses, as well as local individual and business informants. This part of the project was conducted in three stages and resulted in defining strata of telephone exchanges corresponding to sets of zip codes in each of the four cities. While the resulting strata did not cover the full geographic area, based on this extensive preliminary effort it is

estimated that the under-coverage bias ranges from 2% to 12% for the four cities (L. Polack, personal communication, 1998). This figure does not reflect the additional under-coverage within sampled areas of gay males living in non-telephone households. This effort was conducted by a research team led by Dr. J. Catania prior to the telephone sample design and selection.

As has been shown (see Hansen, Hurwitz, and Madow, 1953), when costs differ substantially, optimum sample allocations (assuming constant population variances across strata) assign sample sizes that are inversely proportional to the square root of stratum unit costs. In other words, one allocates more of the fixed sample size to those strata where the cost per case is lower. The strata sample sizes were based on the expected cost per case in each stratum. Initially, four strata were constructed in San Francisco, Los Angeles, and New York, with three strata constructed in Chicago. It was expected that the hit rate for self-identified gay male households, and hence the cost to locate and interview a case, would vary considerably within each city. The distribution of expected hit rates by telephone exchange area was examined for each city and the strata boundaries were set at natural break points in that distribution. There were only two natural breaks for Chicago, so fewer strata were constructed there.

It was expected that there would be differences between the

estimated hit rates and those actually realized in telephone screening. First, there were some inaccuracies in the data sets used for estimation; second, the definition of the survey population did not exactly match that of the various data sources, though there was obviously a close relationship; and third, self-identification of gay status in a telephone survey may well differ from figures obtained in other circumstances.

In phase one, it was planned to select list-assisted samples in each stratum. In phase two, in some strata the identified first-phase numbers would be used to define Mitofsky-Waksberg clusters (banks), in which screening would be conducted until a predetermined cluster size was attained (i.e., a specified number of gay male households was identified).

ADAPTATIONS

Data collection in the cities was begun in waves, so that some information from one city was available when data collection was just starting in another. The order of city data collection was San Francisco, New York, Los Angeles, and Chicago. This proved useful since the total budget was fixed; if the data collection costs in one city appeared to be different from what was anticipated, plans in another city might be affected. The sequential nature of the data collection permitted estimation of costs before fully committing to a sample size in another city.

The final sample size was, in fact, smaller than originally planned due to costs. An unanticipated advantage of the two-phase design was to provide data from phase one, thus breaking the study into manageable components so that changes in sample size could be decided early and that other decisions (such as redefining stratum three in New York) could be made before inordinate amounts of screening were done. Had this not been the case, the final sample size would have been even smaller, perhaps compromising key analysis needs.

In some strata the cost for the intended yield was much higher than anticipated. In those cases, the strata below them, where the yield was expected to be even lower, were dropped. This added minimally to undercoverage, but was cost effective. It is important to note that estimates, strictly speaking, can not be extended to gays living in excluded areas of the cities. This strategy eliminated one stratum each in New York, Los Angeles, and Chicago. In dropping these strata, those resources could be re-allocated to relatively more productive areas and to those areas where the cost per case was turning out to be only somewhat higher than expected. Within stratum three in New York, it was found that some exchanges were considerably more productive than others. The third stratum was split into high and low yield segments, with sampling continuing in the high yield portion and stopped in the other. This strategy, suggested

by Kalton (personal communication, 1977), proved to be a very effective adaptation. Finally, the original cluster size had to be scaled back. This was done for two reasons. First, the overall sample size was reduced, and second, the larger cluster sizes were not attainable without increasing the data collection period, which would have affected costs further. However, as is shown below, the reduced cluster size had the advantage of reducing the variation in the size of the weights.

An examination of Table 1 shows that overall the two-phase design was effective. It was anticipated that gains of 20% would be realized between phase one and phase two. That is, the hit rate using the Mitofsky-Waksberg banks identified in phase one would be 20% higher than the simple RDD rate. As can be seen, the actual gains were more modest. Still, useful gains in yield of self-identified gay males were realized in San Francisco with increases of 5.6% and 21.6% between phase one and phase two and also in New York (31.9% and 13.6%). There were slight increases in Los Angeles (8% and 1.3%) and no increase in Chicago (-3%). This suggests that cities with larger, contiguous gay areas also have moderate clusters of gays outside those known neighborhoods. It is also important to note that, with the exception of San Francisco, the phase one hit rates were uniformly lower than anticipated from the prior analysis of secondary sources. In New York, for example, the actual phase one rates were only slightly

more than half of the expected rates. Without the substantial gains resulting from the adaptations and the second-phase cluster design payoff, the cost per case of a simple screening design would have been about twice the anticipated cost.

Table 1. Expected vs. Actual Hit Rates (%)

	Expected Hit Rates		Actual Hit Rates	
	Phase 1	Phase 2	Phase 1	Phase 2
San Francisco				
Stratum 1	17.2	20.6	28.5	30.1
Stratum 2	11.5	13.8	11.1	13.5
Stratum 3	4.1		9.4	
Stratum 4	1.4		7.9	
Chicago				
Stratum 1	9.9	11.9	6.9	6.7
Stratum 2	6.6		5.5	
Los Angeles				
Stratum 1	23.1	27.7	16.3	17.6
Stratum 2	8.7	10.4	8.0	8.1
Stratum 3	5.7		5.6	
New York				
Stratum 1	20.2	24.2	11.6	15.3
Stratum 2	12.0	14.4	6.6	7.5
Stratum 3	6.1		4.6	

The issue of cluster size is more complex than expected, and more problematic. The cluster size was not reached in the majority of the banks where the standard Mitofsky-Waksberg replacement procedures were used. The main difficulty, which was not anticipated, is that when household eligibility cannot be determined because of non-contact, the number cannot be replaced. In the case of a refusal to be screened the number can be replaced, since the population definition is limited to self-identified gay males. Operationally, that has to mean self-identification in the telephone screener; however, since the target, k, was small the variation in cluster size is not large.

Table 1 also demonstrates that even though extensive and careful efforts were devoted to constructing estimated hit rates based on secondary data, these rates were often not accurate. This finding in itself supports the continued exploration of two-phase adaptive samples in other studies with similar requirements when population-frame parameters are unknown or of suspect reliability.

DISCUSSION

The sample design used in this study can be very effective in some circumstances, assuming the increased sampling variances are not too great. In this design when cluster sizes vary, so do the probabilities of selection. When the probabilities of selection differ for subsets of the sample, the resulting

population parameter estimates will be biased unless weights are used. When the size of weights varies greatly among sample subsets, typically the sampling variances increase. In weighting data for estimation of population parameters the objective is to produce unbiased (or nearly unbiased) estimates without substantially increasing sampling variances.

Waksberg (1993) has shown, following Kish (1965), that for this type of cluster design the increase, D, in variance over the variance of a simple random sample of the same size can be modeled simply. Clusters are first combined into sets, one set for each unique weight. Then $D = [\sum W_j P_j] [\sum W_j / P_j]$ where W_j is the weight for the $_j$th set of clusters and P_j is the percentage of all sample households in $_j$. Waksberg (1993) showed that these increases can be substantial and must be considered against the reduced cost or conversely, the increased yield for the same cost, under this design. However, Waksberg's analysis does not consider a commonly used strategy to control the effect of weighting, which is to trim the weights. In this procedure, extreme weights are reduced somewhat in magnitude: some increase in the bias of the estimator is accepted to prevent large increases in variance. The value D can be computed with the standard weights and then with the trimmed weights. The amount of trimming can be adjusted based on the magnitude of D that is considered tolerable. This is an iterative process whose goal is

to balance sampling variance and bias. The target cluster sizes in San Francisco and New York were seldom achieved, so considerable use of weights was necessary. However, the weights can be based on the achieved modal rather than target cluster sizes. This approach, used in conjunction with weight trimming, can be applied to minimize the variation in weights and, therefore, can control the increase in sampling variances. If the range of cluster sizes is small, the variation in the size of the weights will be small as well; for example, an examination of the range of cluster sizes in San Francisco shows that of the 162 clusters, the modal cluster size was 2. All but 12 cluster sizes are in the range of 1-3. In addition, 11 clusters were size 4 and one was size 5[3]. The size of the weights and their variation in the size can thus be acceptably controlled.

Finally, the adaptations have statistical implications. Where interviewing was curtailed in very unproductive strata, the undercoverage bias increases; however, all available evidence suggests that the increase is small. It should be noted that, as in any cluster design, there is also an effect on the sampling variance due to intracluster homogeneity -- that is, the tendency for observations within a bank to be correlated. The cluster sizes used in the study, however, are so small that effect is quite minor.

One area not sufficiently examined in the reported study is

undercoverage bias. This bias results mainly from the exclusion of some areas of each city from the sampling frame because the estimated cost per case was extremely high. The direction and magnitude of this undercoverage bias could be obtained from a sample of a subset of the excluded areas, though the costs would be high. Undercoverage bias (whether from exclusion of geographic areas, nonresponse, or false negatives) can be modeled as:

$$\text{Bias} = P[E(X_i) - E(X_e)]$$

where P is the proportion of the population not covered, $E(X_i)$ is the expected value of the mean for the included population, and $E(X_e)$is the expected value of the mean of the excluded population for some study variable. Bias increases when there is an increase in either the proportion of the population not covered or in the difference between the value of a variable for the covered and non-covered population components. Bias is zero either when there is no undercoverage or when there is no difference between the population components.

While not totally successful, this design did achieve two key objectives. First, it increased the yield of the target group while maintaining a cost-effective probability sample design. Second, the various adaptations were essential to successfully dealing with the differences between anticipated hit rates and actual hit rates. For these reasons, the design

deserves further testing on different populations. There are two additional approaches to the cluster size problem. First, when the population is relatively rare there is some justification for replacing unknown-eligibility cases, on the assumption that they are likely to be ineligible. Second, projections such as those done in the current study can be used. The latter solution is less desirable since it is likely to exacerbate effect of the weights on sample variances.

In summary, while the design was effective in its key aspects there are problems that need further methodological research to address. Additionally, some unexpected problems arose, but ultimately the sample design approach helped achieve a probability design under very difficult circumstances. In the planning phase it should be recognized that estimated prevalence is being modeled and that alternative models should be compared. Records should be kept comparing yields from surveys to those anticipated from individual or combined secondary sources. Such data could be useful in making more accurate prevalence estimates in future studies.

It is possible to select probability samples and to estimate their undercoverage bias; however, this is a very expensive and time consuming process. Often conducting careful probability surveys (whether or not they include special studies to estimate nonresponse) may be beyond the resources of individual

researchers. One solution may be for researchers to give consideration to collaborating to pool resources to conduct large, multi-purpose surveys of gay males. Once such a sample is obtained, it might be possible (for research purposes where panel effects are not a concern) to maintain the sample as a panel. Sampling gay males and other groups at high risk for HIV is an important area of survey design. To reduce the prevailing reliance on convenience sampling, new ideas for both prevalence estimates and sample designs must continue to be tested.

Footnotes

1. The researchers who conducted this work, and were responsible for the overall GUMS (Gay Urban Males Survey), defined the target groups as MSM's (Men who have Sex with Men). The definition includes men who do not identify themselves as gay, homosexual or bisexual. The present paper uses the term gay for convenience.

2. The study also oversampled high income households in a separate sample using the same methods. Similar gains in yield were obtained.

3. The target cluster size can be exceeded if sample is released in a cluster based on an average projected hit rate. Such projections were used for some strata to speed up the data collection.

References

Binson, D., Moskowits, J., Mills, T., Anderson, K., Paul, J., Stall, R. & Catania, J. (1996). Sampling men who have sex with men: Strategies for a telephone survey in urban areas in the United States. Proceedings of the Section on Survey Research Methods, American Statistical Association, USA, 68-72.

Blair, J. & Czaja, R. (1982). Locating a special population using random digit dialing. Public Opinion Quarterly, Vol. 46, 585-590.

Hansen, M. H., Hurwitz, W. N. W. & Madow, W. G. (1953). Sample survey methods and theory, Vol. 1. New York: Wiley.

Inglis, K. M., Groves, R. M. & Heeringa, S. G. (1987). Telephone sample designs for the U.S. black household population. Survey Methodology, Vol. 13, 1-14.

Kish, L. (1965). Survey sampling. New York: Wiley.

Neyman, J. (1934). On the two different aspects of the representative method. Journal of the Royal Statistical Society, 97, 558-625.

Sudman, S., Sirken, M. G. & Cowan, C. D. (1988). Sampling rare and elusive populations. Science, 240, 991-996.

Sudman, S. (1985). Efficient screening methods for the sampling of geographically clustered special populations. Journal of Marketing Research, Vol. 22, 20-29.

Thompson, S. K. (1992). Sampling. Wiley.

Waksberg, J. (1983). A note on locating a special population using Random Digit Dialing. Public Opinion Quarterly, Vol. 47, 576-578.

Waksberg, J. (1978). Sampling methods for Random Digit Dialing. Journal of the American Statistical Association, Vol. 73, 40-46.