PRACTICAL METHODS FOR SAMPLING RARE AND MOBILE POPULATIONS

Graham Kalton
Westat, 1650 Research Boulevard, Rockville, Maryland 20850

Key Words: Disproportionate Stratification, Screening with Area Sampling, Location Sampling

1. Introduction

Seymour Sudman was an internationally renowned survey methodologist who wrote widely on many aspects of survey methodology. This paper, which was presented in the Seymour Sudman Memorial Session, reviews an area of survey sampling in which Seymour made many important contributions. Seymour's work in survey sampling was firmly grounded in his practical experience. This experience is clearly to be seen in his papers on the frequently encountered and challenging problems involved in devising efficient and valid methods for sampling rare and mobile populations.

Many surveys focus on a subset of the total population, and that subset is often a small proportion of the total. Thus, for example, surveys may be concerned with minority populations, specific age/sex groups such as males aged 18 to 24, the disabled population, or persons with rare diseases. Sometimes a separate sampling frame with good coverage of the rare population is available, in which case the rare population can be sampled using standard methods. For example, a sample of births in 2001 is being drawn from birth certificate records for the National Center for Education Statistics' Early Childhood Longitudinal Study – Birth Cohort. However, in most cases such frames are unavailable, and special sampling techniques are required.

A related topic is sampling mobile populations, such as international travelers, car passengers, visitors to museums or national parks, the homeless, voters at polling booths, hospital outpatients, and shoppers at a shopping mall. Sometimes methods for sampling mobile populations are used for sampling rare populations as discussed later.

Two main objectives in surveys of rare and mobile populations can be distinguished. One is simply to estimate the number of members of the rare population M and the prevalence of the rare population in the total population $P = M/N$, where N is the size of the total population. The second objective is to estimate characteristics of the rare or mobile population, such as the mean for some variable y, $\bar{Y} = \sum Y_i / M$. The proportion of the rare or mobile population with a given characteristic can also be represented by $\bar{Y}$, with $Y_i = 1$ if individual i has the characteristic and $Y_i = 0$ if not. This review is restricted to the second of these objectives.

A wide variety of methods has been used for sampling rare and mobile populations, including:

- Special lists;
- Multiple frames;
- Screening;
- Disproportionate stratification;
- Multiplicity sampling;
- Snowballing;
- Adaptive sampling;
- Multipurpose surveys;
- Location sampling;
- Cumulating cases over several surveys; and
- Sequential sampling.

Methods for sampling rare populations have been reviewed by Sudman and Kalton (1986), Sudman, Sirken and Cowan (1988), Kish (1965a, 1991), Kalton and Anderson (1986), and Kalton (1993a). The use of adaptive sampling for estimating the size of a rare population is not covered in these papers, being of more recent origin. It is described in Thompson and Seber (1996). Kalton (1991) reviews methods for sampling mobile populations. This paper cannot attempt a complete review of the methods. Instead, it will be confined to three widely used techniques: disproportionate stratification, screening in the context of area sampling, and location sampling.

2. Disproportionate Stratification

Some rare populations are more heavily concentrated in certain parts of the population. When this concentration occurs, it can be advantageous to sample the parts with the heavier concentrations at high rates. Thus the various parts of the population are treated as strata, with higher sampling fractions being used in the strata with the greater concentrations. This procedure can be cost efficient since less screening is needed in the strata with higher concentrations to identify members of the rare population.

Assume, for simplicity, the following:

- The population mean of variable y in stratum h ($\bar{Y}_h$) is the same for all strata, i.e. $\bar{Y}_h = \bar{Y}$ for all h.

- The element variance of y in stratum h, S_h^2, is the same for all strata, so that $S_h^2 = S^2$.
- The cost of screening out a member of the non-rare population is the same for all strata, $c_{Sh} = c_S$.
- The cost of collecting data for a member of the rare population is the same for all strata, $c_{Rh} = c_R$.

Then, with simple random sampling within the strata, the optimum sampling fraction in stratum h is

$$f_h \propto \sqrt{\frac{P_h}{P_h(r-1)+1}} \qquad (1)$$

where $P_h = M_h / N_h$ is the prevalence of the rare population in stratum h and $r = c_R/c_S$ (Kalton, 1993a). In most cases, r is greater than 1, often appreciably so. However, when the screening costs dominate, r may be approximately 1. With $r = 1$, the optimum sampling fraction reduces to $f_h \propto \sqrt{P_h}$. Consider, for example, two strata, with 64 percent of the members in one stratum and 4 percent in the other being members of the rare population. Then, with $r = 1$, the first stratum should be sampled at a rate 4 times as large as the second stratum. With $r = 4$, the first stratum should be sampled at a rate only 2.48 times larger than the second stratum.

The gain in precision from the use of disproportionate sampling with optimum sampling fractions over the use of proportionate stratification for the case when $r = 1$ is approximately

$$R = \frac{V(\bar{y}_{opt})}{V(\bar{y}_{prop})} = (\Sigma \sqrt{A_h W_h})^2 \qquad (2)$$

where A_h and W_h are the proportions of the rare and of the total population in stratum h. This formula clearly illustrates the need for the distributions of the rare population and of the total population to differ across the strata if a reduction in variance is to be obtained from a disproportionate allocation.

In the case of two strata, the above formula may be re-expressed in terms of the A_1 and P_1/P, where P_1 is the prevalence of the rare population in stratum 1 and P is its prevalence in the total population. Table 1 presents the value of R (in percent) for different values of A_1 (in percent) and the relative prevalence $RP = P_1/P$. The table shows that the gains from disproportionate stratification are modest (i.e., R is not much less than 1) unless two conditions both apply:

- The prevalence of the rare population in stratum 1 must be much higher than in the total population, i.e., RP must be much greater than 1.
- The proportion of the rare population in stratum 1 must be high, i.e., A_1 must be large.

A value of R of 80 percent is equivalent to a 25 percent variance reduction from the use of the optimum sampling fractions. The stepped line in the table divides the cell values into those with $R > 80$ percent from those with $R < 80$ percent. As can be seen from the table, even when the prevalence of the rare population in stratum 1 is 20 times higher than the average prevalence, a 25 percent or greater reduction in variance is not achieved unless stratum 1 contains more than 30 percent of the rare population. On the other hand, a 25 percent or greater reduction in variance can be achieved if the prevalence of the rare population in stratum 1 is only twice as great as the average prevalence provided that stratum 1 contains 90 percent or more of the rare population.

The reductions in variance that accrue when r is greater than 1 are less than when $r = 1$, in line with the lesser variation in the optimum sampling fractions that occurs in this case, as noted above. Table 2 presents comparable results to those in Table 1, but with $r = 7$. In this case the results depend on the overall prevalence level, which is here taken to be $P = 10$ percent. A comparison of the cell values shows that the R values for given RP and A_1 are appreciably larger in Table 2, and the stepped $R = 80$ line has shifted much further down in the right hand corner of the table. Reductions of variance of 25 percent or more ($R < 0.8$) occur only for values of A_1 of 80 percent or more and then only for sizeable values of RP.

These results show that disproportionate stratification can be useful but that major benefits from the use of this technique arise only when A_1 and RP are large. The benefits decline as the relative cost of data collection from a member of the rare population to the cost of screening out a member of the rare population (r) increases.

Table 1. Values of R for two strata with $r = 1$

	A_1 (Percent)									
RP	10	20	30	40	50	60	70	80	90	100
1	100	100	100	100	100	100	100	100	100	100
1.5	100	99	99	98	97	96	94	92	87	67
2	99	98	97	95	93	91	88	83	76	50
3	98	96	93	91	87	83	78	71	61	33
5	97	93	89	85	80	74	67	59	47	20
10	95	90	84	78	72	64	56	47	34	10
15	94	88	82	75	68	60	51	41	29	7
20	94	87	81	73	66	57	48	38	26	5

Table 2. Values of R for two strata with $r = 7$, $P = 10\%$

	A_1 (Percent)									
RP	10	20	30	40	50	60	70	80	90	100
1	100	100	100	100	100	100	100	100	100	100
1.5	100	100	100	99	99	98	98	97	94	79
2	100	99	99	98	97	96	95	93	89	69
3	99	99	98	97	95	94	91	88	82	58
5	99	98	97	95	93	91	88	83	76	50
10	99	97	96	94	91	88	84	79	71	44
15	99	97	95	93	91	87	83	78	69	42
20	99	97	95	93	90	87	83	77	68	41

3. Screening with Area Sampling

Most national household surveys employ area sampling, and this is also generally so for national surveys of rare populations. Three situations may be usefully distinguished:

1. The rare population is evenly spread throughout the population.
2. The rare population is unevenly spread, with higher concentrations in some areas.
3. The rare population is unevenly spread, with many areas containing no members of the rare population.

These three situations are described in turn below. The section then concludes with a brief discussion of the noncoverage problem that often arises with screening.

3.1 Evenly-Spread Rare Populations

Consider first the estimation of a sample mean $\bar{y}$ from a simple two-stage sample with a equal-sized primary sampling units (PSUs) selected by simple random sampling and b individuals selected by simple random sampling within selected PSUs. Further assume a simple cost model of the form $C = aC_a + abc$, where C_a is the cost of including a PSU in the sample and c is the survey cost per selected individual. Then, from standard theory, the optimum value for b is

$$b_T = \sqrt{\frac{C_a}{c}\frac{1-\rho}{\rho}} \tag{3}$$

where ρ is the intra-class correlation of the *y*-values in the PSUs, and the subscript T denotes the fact that this result relates to an estimate of the mean for the total population (see, for instance, Kish, 1965b).

The above formula for the optimum b can also be applied for an evenly spread rare population (e.g., children aged 18-36 months in an immunization survey), but with the value of c changed to $c' = c_R + c_S(P^{-1} - 1)$, with c_R and c_S as defined earlier. This change applies because the cost of collecting data for a member of the rare population also includes the cost of screening out $(P^{-1} - 1)$ members of the non-rare population. Thus if the cost of survey data collection is the same for a member of the total population as for a member of the rare population excluding the screening cost (i.e., $c = c_R$), then the

optimum value of b_R for the rare population will be smaller than b_T. In this situation, the relationship between b_R and b_T depends on the cost ratio c_R / c_S. If $c_S = 0$, $b_R = b_T$. If $c_R = c_S$, $c' = P^{-1}c_R$, and hence $b_R = \sqrt{P}b_T$. Thus

$$\sqrt{P}b_T < b_R < b_T \qquad (4)$$

It should be noted that, although b_R will generally be less than b_T, the optimum screening sample size per PSU will still be large. For example, with $P = 10$ percent and $b_T = 20$, $6 < b_R < 20$ from the above equation. However, the PSU screening sample size needed to generate such a sample size for the rare population is $63 < n_T < 200$, where n_T is the screening sample size per PSU. Thus, the above discussion does not contradict the well-established advice to select large subsamples from selected clusters when sampling a rare population. However, the optimum subsample size is not as large as would occur by simply equating b_T and b_R.

3.2 Unevenly Spread Rare Populations

Some rare populations are more heavily concentrated in certain areas. For selecting a sample of such a population, disproportionate stratification may be employed, with the strata being defined geographically. As discussed in Section 2, such disproportionate stratification gives notable gains in precision only when the prevalence of the rare population is much higher than average in some geographical strata *and* when these strata contain a substantial proportion of the rare population. Also, as noted in Section 2, the relative cost of a full interview to a screening interview (r) affects the effectiveness of disproportionate stratification. The higher the value of r, other things being equal, the less the gain in precision.

Waksberg, Judkins and Massey (1997) provide an extensive and informative evaluation of the effectiveness of disproportionate geographic stratification for sampling racial minorities and the low income population, where the geographical areas were 1980 and 1990 Census blocks and block groups. Their findings indicate that disproportionate stratification is useful for sampling blacks and Hispanics for $r < 5$ or 10, and for other minorities for even larger values of r. However, the gains from disproportionate stratification for sampling the low income population are small because, although there are areas with high concentrations of the low income population, a high proportion of the low income population lives elsewhere. Waksberg *et al.* also point out that an assessment of the effects of disproportionate geographic stratification based on Census data needs to take into account the changes that will have occurred in the geographic distribution of the rare population between the time of the Census and the time of the survey.

3.3 Many Clusters Containing No Members of the Rare Population

There are some rare populations that go unrepresented in many geographic clusters. If the zero clusters can be identified in advance of the survey, they can simply be removed from the sampling frame. However, when they cannot be identified in advance, under standard designs the zero clusters are sampled and extensive, but unproductive, screening is conducted within them. Sudman (1972, 1985) has proposed a scheme to avoid this outcome. His scheme is based on the Mitofsky-Waksberg random digit dialing scheme for telephone surveys (Waksberg, 1978). The scheme involves the initial selection of one (or a few) elements in each sampled area. If the selected element is a member of the rare population, further screening is carried out until b more elements of the rare population are sampled. If the initially selected element is not a member of the rare population, no more screening is conducted in that area. There are, of course, fieldwork issues to be considered in applying this scheme. However, in some circumstances the scheme can be effective for very rare populations and when there are many zero clusters.

3.4 Noncoverage with Screening

Screening involves collecting data from the members of the initial sample in order to be able to classify them as members or nonmembers of the rare population. The identification of some rare populations requires only one or a few questions (e.g., children aged 18-36 months), but for other rare populations many questions may be needed (e.g., low-income white families with a male head under 25 and 2 or more children). Misclassification errors at the screener stage can give rise to serious levels of noncoverage (Sudman, 1972, 1976). Misclassifications of nonmembers as members (false positives) are usually corrected in the detailed interview that follows the screener, but misclassifications of members as nonmembers (false negatives) are not corrected and thus result in noncoverage. The risk of false negatives is heightened when the screener respondents can deduce the rare population of interest from the contents of the screener questions or from advanced material supplied to them, since they can then avoid the full interview through their choice of responses to the screener. Thus, designers of screening questionnaires generally attempt

to avoid a transparent disclosure of the rare population. Even so, substantial levels of noncoverage of rare populations are widely encountered with screening, and particularly so when many questions need to be used to identify the rare population, and an incorrect answer to any one of them leads to a false negative outcome. As a typical example, Horrigan *et al.* (1999) report that only 75 percent of youths aged 12-23 years of age were located in the National Longitudinal Survey of Youth of 1997 (NSLY97); much of the loss was probably due to noncoverage.

Another issue is nonresponse at the screening phase. There is often a concern that screener nonresponse will be higher for the rare population than the total population. Thus, even a high overall screener response rate may mask a low response rate for members of the rare population.

4. Location Sampling

Location sampling refers to methods used to sample individuals who visit specific locations such as libraries, museums, shopping centers, and polling places. Sampling is usually conducted either as the visitors enter or as they leave a location. Two distinct units of analysis need to be distinguished – visits and visitors (Kalton, 1991). Location sampling can readily produce a probability sample of visits, with known selection probabilities, and hence visits are easily analyzed. Visits may be the appropriate unit of analysis for, say, a survey about satisfaction with visits to a museum. However, for many surveys using location sampling, the visitor is the appropriate unit of analysis. For example, the visitor is the appropriate unit of analysis in a survey of visitors to soup kitchens over a week to estimate the number of homeless, a survey of nomads visiting watering holes to estimate the size of the nomadic population, or a survey of men who have sex with men (MSM) visiting gay bars to study the characteristics of the MSM population.

The use of the visitor as the unit of analysis is complicated by the fact that a visitor may make multiple visits during the survey's time frame. If a standard sample of visits is selected, the increased selection probabilities associated with multiple visits need to be taken into account in developing the survey weights. The problem lies in estimating the multiplicities, both because a sampled person may be unable to accurately recall past visits since the start of the survey's reference period and because he or she is unable to forecast visits to be made from the time of interview until the end of the reference period. As a result, the multiplicities may be based on simple reports about general frequency of visits.

An alternative solution to the multiplicity problem is to uniquely identify one of the visits with the visitor, treating the other visits as blanks, thereby avoiding the problem. The natural choice for the uniquely identified visit is the first one in the survey reference period: each sampled person is asked if the visit is his or her first since the start of the survey, is selected if the answer is "Yes", and is rejected if the answer is "No". From the fieldwork perspective, an unattractive feature of this procedure is that most visits near the start of the time period will be first visits, leading to interviews, whereas most near the end will not. To some extent, this problem can be addressed by sampling the time periods with probabilities proportional to appropriate size measures, but determining these measures is problematic.

The usual sample design for a location sample is a two-stage design (Kalton, 1991). Primary sampling units are constructed as combinations of locations (entrances or exits) and time segments when the location is open (e.g., a given Monday from 10 a.m. to 2 p.m.). The PSUs are sampled with probabilities proportional to size, with careful stratification by location and time. Then some form of systematic sample is employed to select visitors entering (or exiting) the location. Sudman (1980) outlines the application of this type of design for sampling visitors to a shopping center with several entrances and using half-hour time segments when the center is open.

Location sampling has been widely used for surveys of MSM concerning HIV risk and illness (Kalton, 1993b) with locations such as bars, dance clubs and street locations where MSM congregate. The Young Men's Survey conducted in 7 cities in 1994-1998 in 194 public locations is a major survey of this type (Valleroy *et al.*, 2000). It had a sample size of 3,492 MSM aged 15 to 22 years of age who consented to an interview and HIV testing. A complex weighting scheme was devised to address the multiplicity problem (MacKellar *et al.*, 1996).

5. Concluding Remarks

The sampling of rare and mobile populations often presents survey statisticians with major challenges. Although many methods have been devised for sampling these populations (only a few of which have been discussed here), finding cost-effective methods is frequently difficult and requires ingenuity. In a number of cases, a compromise needs to be made between scientific rigor and practicability. When this occurs, a careful assessment of likely biases and good judgment are required. Seymour Sudman's many valuable contributions to the subject exhibit this combination of ingenuity with a thoughtful balance of practicability and scientific rigor.

6. References

Horrigan, M., Moore, W., Pedlow, S., and Wolter, K. (1999). Undercoverage in a large screening survey of youths. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 570-575.

Kalton, G. (1991). Sampling flows of mobile human populations. *Survey Methodology*, 17, 183-194.

Kalton, G. (1993a). *Sampling Rare and Elusive Populations*. New York: Department of Economic and Social Information and Policy Analysis Statistics Division, United Nations.

Kalton, G. (1993b). Sampling considerations in research on HIV risk and illness. In *Methodological Issues in AIDS Behavioral Research*, D.G. Ostrow and R.C. Kessler eds., pp. 53-74. New York: Plenum Press.

Kalton, G. and Anderson, D.W. (1986). Sampling rare populations. *Journal of the Royal Statistical Society, A*, 149, 65-82.

Kish, L. (1965a). Selection techniques for rare traits. *Genetics and the Epidemiology of Chronic Diseases*, Public Health Service Publication No. 1163.

Kish, L. (1965b). *Survey Sampling*. New York: Wiley.

Kish, L. (1991). Taxonomy of elusive populations. *Journal of Official Statistics*, 7, 339-347.

MacKellar, D., Valleroy L., Karon, J., Lemp, G. and Janssen, R. (1996). The Young Men's Survey: methods for estimating HIV seroprevalence and risk factors among young men who have sex with men. *Public Health Reports*, 111(Suppl. 1), 138-144.

Sudman, S. (1972). On sampling of very rare human populations. *Journal of the American Statistical Association*, 67, 335-339.

Sudman, S. (1976). *Applied Sampling*. New York: Academic Press.

Sudman, S. (1980). Improving the quality of shopping center sampling. *Journal of Marketing Research*,17, 423-431.

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

Sudman, S. and Kalton, G. (1986). New developments in the sampling of special populations. *Annual Review of Sociology*, 12, 401-429.

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

Thompson, S.K. and Seber, G.A.F. (1996). *Adaptive Sampling*. New York: Wiley.

Valleroy, L.A., MacKellar, D., Karon, J. *et al.* (2000). HIV prevalence and associated risks in young men who have sex with men. *Journal of the American Medical Association*, 284, 198-204.

Waksberg, J. (1978). Sampling methods for random digit dialing. *Journal of the American Statistical Association*, 73, 40-46.

Waksberg, J., Judkins, D. and Massey, J.T. (1997). Geographic-based oversampling in demographic surveys of the United States. *Survey Methodology*, 23, 61-71.