



Catalogue no. 11-522-XIE

**Statistics Canada International Symposium  
Series - Proceedings**

**Symposium 2004: Innovative  
Methods for Surveying  
Difficult-to-reach Populations**

2004



## APPLICATIONS OF ADAPTIVE SAMPLING PROCEDURES TO PROBLEMS IN PUBLIC HEALTH

Myron J. Katzoff<sup>1</sup>

### ABSTRACT

Adaptive sampling procedures have significant potential value for application to emerging public health problems. It is anticipated that this will especially be the case when dealing with the consequences of bioterrorism events. Since such events are likely to initially impact only a small portion of a human population, the few infected individuals and persons with whom they have had contact may be difficult-to-locate. Adaptive methods are designed to take advantage of relationships which link sample elements to each other and, therefore, can be expected to yield samples containing more of the important sample elements than would be obtained from conventional sampling methods. This paper will examine how adaptive sampling methods might be used to extend current national health surveys to enable effective tracking and monitoring of new forms of health threats and trace exposed persons.

KEYWORDS: Adaptive Sampling; Network Sampling; Random Walk Designs

### 1. INTRODUCTION

Adaptive designs are an important class of survey design options for:

- (1) ensuring adequate numbers of the sample units of interest for reliable estimation of the characteristics for elusive or hard-to-locate population subdomains;
- (2) achieving suitable sample-content enrichment (for example, with regard to the ranges of explanatory variables) for secondary analyses of survey data or for detailed problem-related study; and
- (3) controlling "case" content (for example, as defined by the numbers of sample units corresponding to various combinations of demographic variables) when there may be concerns about the quality and relevance of a sample selected for general population estimates.

Since disease outbreaks and bioterrorism events are likely to initially impact only a small portion of a human population, the few infected individuals and persons with whom they have had contact may be difficult to locate. Thus, adaptive sampling procedures can potentially improve the quality of statistical information for dealing with the consequences of these public health challenges.

When subdomain sample sizes are dynamically controlled by adaptive sampling procedures, data collection instruments can provide information in sufficient quantity to track the effectiveness of interventions over time and to identify how and where a disease is spreading. This might be accomplished very effectively by adaptively resurveying the affected population groups at regular intervals. Additional uses of data obtained from the application of adaptive procedures include monitoring and forecasting the numbers of individuals in the various stages of a disease (via, for example, state-space models) and updating and validating mathematical models used in the analysis of the population dynamics of epidemics.

The feature of adaptive designs which distinguishes them from conventional designs is that the inclusion of units in the sample is allowed to depend on the values of variables observed for sample units chosen during the survey. This feature has important consequences for statistical inference from data obtained from adaptive sampling procedures. Units added adaptively to the sample by following links can result in samples with a probability structure which is

---

<sup>1</sup>Myron J. Katzoff, National Center for Health Statistics, Centers for Disease Control and Prevention, 3311 Toledo Road, Hyattsville MD, USA, 20782 (mkatzoff@cdc.gov)

markedly different from that of a conventional sampling procedure. The estimators for conventional designs cannot be used without adjusting in some manner for changes in the probability structure introduced by following links. To provide some idea of what makes a sampling procedure “adaptive” and how the probability of inclusion for a unit could be affected, it might be useful to keep in mind that some basic adaptive designs involve the selection of an initial sample according to a conventional procedure; units are then added to that sample in conformance with a rule which directs that further selections be made (or not made) for each element of the initial set of units, and other units already in the sample, depending upon the information observed for each unit. Additional details on these designs can be found in Thompson (1992).

As used here, “adaptive sampling design” covers link-tracing designs such as network sampling, snowball sampling, chain-referral sampling, respondent driven sampling, random walk designs and adaptive cluster sampling. It also includes active set adaptive designs, the newest class of designs (which includes those just listed as special cases), in which units are added to a sample according to a mixture of two distributions: with high probability, the units selected in a wave of sampling are determined by the values of variables measured for some subset of the units already in the sample and identified as the “active set”; and, with lower probability, a selection is made based on the sampling frame structure without reference to the values of any observed variables.

**Acknowledgement:** I have drawn heavily from the concepts and ideas which can be found in the works of Monroe G. Sirken and Steven K. Thompson. I have also benefited from numerous discussions with these distinguished researchers.

## 2. A FEW FREQUENTLY USED TERMS

In some discussions of adaptive designs, the authors have found it convenient to use the terminology and concepts of graph theory to describe populations with linkage or networks. When graph theory diagrams are being used, the nodes of graphs represent the population elements of interest and the edges (sometimes called arcs) connecting units represent the links between units. In addition, depending upon the application, it may be important to distinguish between observational units, the individual units of a population under study (hence, the nodes), and selection units, the members of a sampling frame or universe employed to acquire a sample of observational units. The observational and selection units may be the same or different.

## 3. SPECIAL FEATURES OF THE NHIS DESIGN FOR ADAPTIVE SAMPLING

There are some special features of the National Health Interview Survey (NHIS) that make it an attractive mechanism to which adaptive sampling procedures might be appended. The illustrative, though speculative, examples presented in later sections exploit these features.

The NHIS is a national stratified multistage survey in which households are the final stage selection units. The following design features would facilitate the application of adaptive sampling methods locally and, possibly, on a national scale if required:

- (1) state is a stratification variable;
- (2) primary sampling units (PSUs) are counties, groups of counties, county equivalents, (e.g., parishes and independent cities), towns, townships, minor civil divisions or metropolitan statistical areas (MSAs);
- (3) some large PSUs are selected with certainty and are, therefore, strata themselves; and
- (4) when necessary and provided that there has been no change in the general survey design for adjacent time periods, (conventional) samples for those time periods can be grouped to ensure large enough initial samples needed for the subsequent adaptive inclusion of observational units to cover 80+ areas designated as consolidated metropolitan statistical areas (CMSAs), primary metropolitan statistical areas (PMSAs) and MSAs.

## 4. A NETWORK SAMPLING EXAMPLE WITH SYMMETRIC LINKS

Suppose the statistical objective is to estimate for each state for the three largest Asian subpopulation subgroups in the nation the number of individuals with a genetic disorder. To accomplish this objective, consider adding individuals to the NHIS conventional household sample by tracing family links. In this example, the adults and children who are members of the population of a state will be the nodes. The arcs will depict links defined by family relationships: parents, parental siblings, children and their siblings, and the children of siblings who are the residents of a state. Assume the survey instrument and interviewing protocol are designed so that the inclusion of any one observation unit of a linked group results implicitly or explicitly in the inclusion of the data for all the linked observation units of that group. The links are then called symmetric. (It should be especially noticeable that the chance inclusion of any node in the domain of interest results in the addition of data on all linked population members to the sample.) Notice that the interviewing protocol might allow for a variety of interviewing modes or substitute respondents which could reduce data collection costs in some applications.

The NHIS employs a complex array of sample-weight adjustments designed to produce approximately unbiased design-based national estimates of population totals. A national-estimates weight pertaining to estimates for persons is the product of four factors:

- (1) the base weight, which is the inverse of the probability of selecting a household;
- (2) a household nonresponse adjustment;
- (3) a first-stage ratio adjustment, to control person-level weights for PSUs not selected with certainty to population totals for 24 residence and race-ethnicity classes; and
- (4) a second-stage ratio adjustment, to control person-level weights to agree with independently determined totals prepared by the U.S. Bureau of the Census for 88 age-sex-race/ethnicity classes for the civilian noninstitutionalized population.

Further details on the purposes and uses of the weight adjustments can be found in Botman, Moore, Moriarity and Parsons (2000). In principle, some modification of the above series of weight adjustments could be applied to obtain estimates for each state. However, one should not expect an effective adjustment-cell scheme to be the same for each state; the effective scheme for each state may depend on the statistical objective and it could involve a great amount of experimentation due to the need to customize the collapsing of adjustment cells for each state.

In order to concentrate on the essential features of the basic design-based estimator for the adaptive sampling procedure discussed in this section, no further consideration will be given to the weight adjustments that very likely would be necessary to address nonresponse and frame problems. Accordingly, one may then regard the main survey for each state as having at most two stages of sampling: (1) counties, or other appropriate areas, as PSUs; and (2) households within PSUs as secondary sampling units (SSUs). For observation unit  $\ell$ , let

$$y_{\ell} = \begin{cases} 1, & \text{if the person is a member of a relevant} \\ & \text{Asian subgroup and has the disorder} \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

If  $A_{ik}$  denotes the set of observation units in the state that are linked to selection unit  $k$  of PSU  $i$  and  $m_{\ell}$  denotes the number of selection units in the state to which observation unit  $\ell$  is linked, define the selection unit average

$$w_{ik} = \sum_{\ell \in A_{ik}} \frac{y_{\ell}}{m_{\ell}} \quad (2)$$

When  $t$  PSUs are chosen without replacement from  $T$  with inclusion probabilities  $\pi_i$  and, within each PSU,  $n_i$  households are chosen from  $N_i$  without replacement under simple random sampling, the two-stage unbiased estimator of the total number of state residents in the target group with the disorder is

$$\hat{\tau} = \sum_{i=1}^t \pi_i^{-1} \frac{N_i}{n_i} \sum_{k=1}^{n_i} w_{ik}. \quad (3)$$

There is a contribution to each estimate corresponding to  $\hat{\tau}$  for each stratum from which PSUs are not chosen with certainty. Since the incidence rate by state could be of greater interest to investigators, one should note that it would be easy to define indicator variables of the form of  $y_{\ell}$  for that purpose which separate the presence or absence of the disorder from race membership. The incidence rate is then, of course, the ratio of two estimates of totals for each state. In that case, the estimators for numerator and denominator would each include terms of the form of  $\hat{\tau}$ .

Straight-forward design-based expressions for the variance of  $\hat{\tau}$  and its estimator are given in Katzoff, Sirken and Thompson (2002). However, in any application that involves a supplementary use of the NHIS, numerous sampling effects which have not already been discussed must be taken into account in the estimation of variances. For practical data analysis, all-purpose variance estimation structures which are easy to implement with existing computer software will be needed. These points are discussed in detail in Botman, Moore, Moriarity and Parsons (2000).

Some final remarks about this example are in order. It may be of interest that a careful reading of the first example in Katzoff, Sirken and Thompson (2002) leads to the realization that there is a subtle difference between the two examples even though the algebraic forms of the estimators are the same. The links in this example may be said to be strongly connected, the term which applies under complete link-tracing when a sample including any one observation unit of a linked group results in adding all the members of that group to the sample. In practice, links may not be defined so that units are strongly connected and complete tracing of the links of a connected component may be impractical because of limited resources.

## 5. AN ADAPTIVE SAMPLING EXAMPLE USING A RANDOM WALK DESIGN

In this example, the statistical objective is to estimate the proportion of a local population, a CMSA, PMSA or MSA, say that has been exposed to an easily transmitted contagious biological agent. The NHIS sample is again used as a starting point. The nodes are individual members of the population and the arcs are determined by mechanisms that would enable contact and transmission among population members. For the application of the random walk procedure described in this section, the influence of the specific features of the NHIS design is diminished as sampling continues; in fact, the initial sample of persons drawn from households may be somewhat uncontrolled. This is due to monitoring and controlling the link tracing procedure to asymptotically yield desired selection probabilities as described further below. The basic idea is to work with the natural tendencies of populations and link-tracing procedures to provide just enough direction during the sampling so that simple estimates can be calculated which are representative of the population as a whole. The approach taken here is due to Thompson (2003).

For the random walk design, a Markov chain structure that is aperiodic and irreducible is needed. Let a fundamental set of transition probabilities be defined by

$$q_{ij} = \begin{cases} \frac{(1-d)}{N} + \frac{d a_{ij}}{a_i}, & \text{if } a_i > 0 \\ \frac{1}{N}, & \text{if } a_i = 0 \end{cases} \quad (4)$$

where  $a_{ii} = 0$ ; for  $i \neq j$ ,  $a_{ij}$  is the number of links between units  $i$  and  $j$ ;  $0 < d < 1$ ; and  $a_i \stackrel{\text{def}}{=} \sum_j a_{ij}$ , the out-degree for node  $i$ . Next, suppose that members of the population are to be selected approximately according to a given stationary distribution  $\{\pi_1, \pi_2, \pi_3, \dots\}$ . The probabilities  $\pi_i$  might be determined as functions of demographic variables. For example, based upon knowledge of individual vulnerability to the agent, the population is divided into two groups and individuals in the first group are selected at a rate that is twice that of the second group. Suppose that, based upon a recent census, it is known that there are  $N_1$  individuals in the first group and  $N_2$  individuals in the second group. In this case, one might aim to consider selecting individuals in the first group with probability  $\frac{2}{2N_1 + N_2}$  and individuals from the second group with probability  $\frac{1}{2N_1 + N_2}$  so that the  $\pi_i$  take on one, and only one, of these values.

In order to guide the random walk toward the stationary distribution  $\{\pi_1, \pi_2, \pi_3, \dots\}$ , the method of Hastings (1970) is invoked. This method imposes upon the walk the transition probability matrix with diagonal elements

$$P_{ij} = q_{ij} \alpha_{ij} \quad (5)$$

where  $\alpha_{ij} = \min\{1, \pi_j q_{ji} / \pi_i q_{ij}\}$ , and diagonal elements  $P_{ii} = 1 - \sum_{i \neq j} P_{ij}$ . The procedure yields data to quantify the generalized ratio estimator of the mean

$$\hat{\mu} = \frac{\sum_{i=1}^n y_i / \pi_i}{\sum_{i=1}^n 1 / \pi_i}, \quad (6)$$

where  $n$  is the sample size with or without repeats. Note that when  $\pi_i = \pi$  for all  $i$ ,  $\hat{\mu}$  is simply the arithmetic mean. Finite simulation studies have so far indicated that the sampling distribution of  $\hat{\mu}$  has a mean that is very nearly that of the known mean for moderate values of  $n$  when the  $\pi_i$  of the stationary distribution have no extreme values. This suggests an important direction for proofs of asymptotic results.

For the estimator of the variance of  $\hat{\mu}$ , Thompson (2003) considered use of the sample permutation distribution conditional on the sufficient statistics. However, this approach can result in negative variance estimates. To avoid that problem, he suggested using an estimator based on a few independent replicates of the random walk.

## 6. FINAL REMARKS

In the first example, it was required that each connected component be completely sampled in conformance with the sampling protocol for unbiased estimation. In the second example, it is hard to determine how long the random walk should be allowed to continue for asymptotic effects to take over. However, the larger class of active set adaptive sampling (ASAS) procedures includes many designs not discussed in this paper that address these problems. They allow for the use of a criterion for adding units which can depend on node and link variables. They also provide flexibility in controlling sample size and which sample units direct the addition of new members of the population to the sample. The estimation procedures for ASAS designs use specific details of the probability structures encountered in each wave of selection when adding units and they exploit ideas like sufficiency and Monte Carlo sampling for unbiased estimation.

Future study of ASAS designs will explore sampling and unit addition mechanisms beyond those discussed in this paper and will include the development of design-based asymptotics in addition to the usual estimators for means and variances. For designs employing an additional sample, a better understanding of the influence of the parameters for drawing the initial sample is needed. This includes, in particular, consideration of the allocation problem: how large should the initial sample be for various designs? Finally, there is a considerable need for a full development of model-based estimation procedures for ASAS designs.

## REFERENCES

- Botman, S. L., Moore, T. F., Moriarity, C. L. and Parsons, V. L. (2000), "Design and Estimation for the National Health Interview Survey, 1995-2004", *Vital Health Statistics, Series 2*, 130, Hyattsville, Maryland: National Center for Health Statistics.
- Hastings, W. K. (1970), "Monte Carlo Sampling Using Markov Chains and Their Applications", *Biometrika*, 57, pp. 97-109.
- Katzoff, M. J., Sirken, M. G. and Thompson, S. K. (2002), "Proposals for Adaptive and Link-Tracing Sampling designs in Health Surveys", *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 1772-1775.
- Thompson, S. K. (1992), *Sampling*, New York: Wiley.
- Thompson, S. K. (2003), "Simulation Program for Link-Tracing Designs: Designs that Take Advantage of inherent Link-Tracing Tendencies and that Avoid Truncation Problems", unpublished report, Hyattsville, Maryland: National Center for Health Statistics.