INVERSE SAMPLING ALGORITHM FOR NHIS CONFIDENTIALITY PROTECTION

Susan Hinkins, Ernst and Young, Van Parsons, National Center for Health Statistics, and Fritz Scheuren, The Urban Institute
Fritz Scheuren, 1402 Ruffner Rd., Alex. VA 22302, <scheuren@aol.com>

Key Words: Analysis of Complex Survey Data

1. Introduction

In many surveys, the finer details of the geographical sampling structures cannot be released to the public because of confidentiality concerns. For example, the National Health Interview Survey (NHIS) uses a state-level stratification and selects counties and metropolitan areas for the sample. If a state database is released, extreme care must be taken to ensure that the user cannot identify smaller geographical areas. If the geographical sampling structures are deleted, then confidentiality may be achieved, but the data become difficult to analyze using standard design-based methods.

Recent work by Hinkins, Oh, and Scheuren (1997) has demonstrated that an inverse sampling algorithm may facilitate the dissemination and analysis of complex survey data. An inverse sample algorithm is a subsampling mechanism on the original sample data that generates a "new" sample that can be treated as a simple random sample from the population. Inverse sampling techniques may provide a useful means to allow the public to have access to micro-level National Center for Health Statistics (NCHS) data because the geographical identifiers would not be needed for the analysis from a simple random sample; a way would exist to estimate variances that would not require geographic data be retained.

This paper is a progress report on how we intend to apply the inverse sample technique to a typical NCHS survey, the National Health Interview Survey. As will be evident, the presentation preserves much of the character of the original talk we gave in Baltimore. In particular, we have retained most of the headings in the transparencies used.

For the talk there were seven slides and this is how we have organized these proceedings. We began by describing the talk's structure (Section 1). Then in Section 2 (Slide 2) we set out the analytic context that motivates our efforts. Section 3 addresses the conflict built into the need to create public use data sets that are valuable to researchers, while still affording the full confidentiality protection promised respondents. This is done specifically in terms of variance estimation for a multistage stratified area probability sample, like that for the NHIS (Section 4). The fifth part of the talk defines the approach we are using -- the particular application of an inverse sampling algorithm (Hinkins, Oh, and Scheuren 1997). How we intend to apply the inverse sampling algorithm approach to the NHIS is still being worked out but in the talk we provided some early ideas (Slide 6). Next steps, and there are lots of these, conclude the paper (Section or slide 7).

2. Analytic Context

On the slide used in the talk, we listed 3 topics that, taken together, defined our analytic context. These were (1) the role of government statistical agencies in free societies; (2) how important public use files are in achieving a statistical agency's role; and (3) the issues around preserving the analytic potential of the publicly available data -- while still preserving the confidentiality promised to respondents. The stage having been set, we ended this slide by applying the context developed to variance estimation in the NHIS.

2.1 Role of Government. We take as given in a free democratic society, like the United States, that the role of a government statistical agency is to maximize the "openness" of its operations -- to the extent that this does not conflict with other equally held values, like keeping sacred the oaths made to respondents to preserve the confidentiality of any data respondents' entrust to the agency. It might be added that there is also an obligation to reduce such conflicts to the greatest extent possible so that confidentiality is not used as a shield against open access to data by organizations or individuals outside government.

2.2 Public Use Files. For federal statistical agencies in the United States, since the early 1960's and the pioneering work of individuals like Jack Beresford and Joe Pechman, public use files have been one of the responses to achieving the goal of "openness." Last year at these meetings, there was a session on the possible need to modify current practice in releasing public use files – to "morph" it. At least some of the speakers at that session (e.g., Mulrow and Scheuren 1998) stressed that this need -- given the Internet, advances in record linkage techniques plus the growing access electronically to data of all sorts, particularly in the private sector.

2.3 Preserving Analytic Potential. How to "morph" or change the data now being released so that it is generally safe is a topic well beyond the goal of this paper. Our task for the NHIS is the more limited one of describing how we intend to protect the analytic potential for calculating survey variances while still meeting the full confidentiality pledge given to NHIS respondents. It is only to this degree that we are going to describe how to do the needed "morphing."

Don Rubin, among others, has advocated that wholly synthetic data be created (Rubin 1993). I believe we should follow his advice when and if it becomes practical. Indeed, we should try to make it practical as some already have (see Kennickell 1999). There are, however, partial solutions (like ours) that are worth considering. Even if our approach can be seen only as a stopgap, it may arguably be a useful stopgap.

2.4 Variance Estimation. The issue in variance estimation on a public use file is that information on the nature of the sample selection must be provided, implicitly or explicitly, for design-based variances to be calculable. This information could be potentially identifying, especially in area probability designs, like the NHIS, where geographically linked information is essential to derive variance estimates.

3. Confidentiality Protection in Area Probability Designs

In the early days of public use file releases, the information needed to calculate variances was often just not provided. Sometimes, though, only obvious identifiers were removed, and deductively the primary sampling units (PSUs) in an area design could still be derived. These were the days, not that long ago, when PSU maps were on many office walls, including those of at least some of the authors of this paper.

3.1 Recent Practice. Many surveys now employ replicates imbedded within the public use file that can be used to calculate variances. The Current Population Survey (CPS), for instance, has been doing this for some time. Effectively, the original geographic and other design information is being transformed into a series of replicates that each has been independently reweighted to estimate the total population.

Consider the National Survey of America's Families (NSAF), for example, which now has released several public use files that allow researchers to calculate design variances using 60 specially weighted half samples (e.g., Leonard, Russell, and Scheuren 1999). These replicates are designed to be used with the generally available software package, Wesvar, to obtain variances. (See also Scheuren 1999.)

3.2 Remaining Weaknesses. The replicate variance approach can still raise confidentiality concerns depending on the details of the underlying sample design. In an area probability design, like the NHIS (which does not now release replicates for variance estimation), the replicates might be chosen from all or a portion of a completely balanced set of half samples (e.g., McCarthy 1968). The replicates can be constructed by treating the design as having two primary sampling units (PSUs) per stratum for non-self representing PSUs. For self-representing PSUs, the secondary sampling units (SSUs) are divided up into the half sample replicates as well (as described in detail in Wolter 1984).

3.3 Need for Further Changes. This would seem to be a good "solution" and it certainly helps greatly in preserving confidentiality. But it is possible to motivate why something more is needed and why an inverse sampling approach may offer another "solution" with conceptually both better confidentiality protection and better variance estimation properties as well – a win-win, in fact.

4. Balanced Half-Samples

When Phil McCarthy, working as a consultant for the National Center of Health Statistics (NCHS), first laid out the theory of completely balanced half-sample replicates in 2 PSU per stratum designs, he made a great advance that all of those who have come after him have benefited from (McCarthy 1968). He, of course, was not trying to solve a confidentiality problem, even though one of the applications of his ideas, as we have seen, does at least partially do this.

4.1 Remaining Confidentiality Problem. The remaining confidentiality problem depends on how many replicates are released on the public use file. Consider the possibility that the full completely balanced set of replicates is released. Then, select any sampled household and ask yourself, over all sets of half samples, what other households are always together with it – i.e., in the same replicate. It is only those in the same PSU. The records so grouped could be summarized and the summaries might first be compared to each other to put the self-representing PSUs back together by combining the SSUs. Then, the PSUs summaries could be compared to information from the previous census, say, to attempt to place the selected observations in a specific locale. If we already

know what state the observations are from, this might not be as hard as it might seem at first. Insider knowledge, like a list of sample PSUs, would, of course, greatly ease this challenge. We conjecture that in some cases at least, it might be possible to re-identify geography details, even without this aid.

4.2 Grouping Half-Samples. Of course, public use files do not generally release the full set of completely balanced half samples. Just a subset of them is released – as, say, in the CPS. Releasing only a random subset (e.g., Gurney 1975) of the completely orthogonal half samples in the original McCarthy design, as seems to be the practice, affords an opportunity for protection of individual PSUs not available with complete balancing. This seems to be the way, for example, that the CPS replicates are set up (Fay, 1999). The replicates chosen are checked to see if the implicit disclosures possible, as described above, reveal geographic units that fall below the CPS disclosure rules. In the CPS these rules state that no geographic area of smaller than 100,000 population can be revealed.

4.3 Degrees of Freedom. Perhaps then there are no problems to be solved here. Just be careful when creating the half sample replicates. Well, this view is legitimate, at least to some extent, although we would argue that the choice of replicates can be messy and the fact that they may no longer be fully random is discomforting.

There is another perspective that bears examination too and which motivated us as well – the loss in efficiency in estimating the variance that the failure to use a fully balanced set of samples entails. In the original paper that first introduced the inverse sampling algorithm (Hinkins, Oh, and Scheuren 1997), the claim was made (and a proof given) that inverse sampling variance estimates could potentially improve on even a completely efficient traditional approach, let alone one that used only a subset of the possible replicates. For a National survey estimate from the NHIS, considerations of efficiency might not be too important but at the state level they could be a major issue, especially in a small state with only a few PSUs.

In the next section (Section 5) we introduce inverse sampling algorithms and in Section 6 sketch how we will apply the algorithm to the NHIS.

5. Inverse Sampling Algorithms

Hinkins, Oh, and Scheuren (1997) introduce a way to invert many complex sample designs so that simple random subsamples are produced. The approach is to resample the complex sample to obtain an easier to analyze data structure. Because any given resample is unlikely to contain all the information in the original survey, the original complex sample is repeatedly resampled.

These "inverse sampling" algorithms, when feasible, make it possible to employ conventional techniques, like regression and contingency table analysis, with only minor adjustments. Taken together such subsamples can be nearly as efficient as the original sample. In this sense, the resulting data sets are "design-free," since the original complexity of the selection process no longer stands in the way of the full use of standard tools. We also believe that, as has been mentioned, they have potential for preserving confidentiality at the same time as they achieve analytic goals.

5.1 Basic Approach. Notice some things that this approach is -- and is not: First, it is extremely computer intensive. Second, it presupposes that practical inverse algorithms exist (which may not always be the case). Third, it also assumes that the original power of the full sample can be captured if enough subsamples are taken, so that no appreciable efficiency is lost. Fourth, as much as it may resemble the bootstrap (Efron, 1979), we are not doing bootstrapping. There is no intent to mimic the original selections, as would be required to use the bootstrap properly (e.g., McCarthy and Snowden, 1985) -- just the opposite; our goal here is to create a totally different and more analytically tractable set of subsamples from the original design.

Suppose that we wish to draw a simple random sample, without replacement, from a finite population of size N. Suppose further that the population is no longer available for sampling, but we have a sample selected from this population using a sample design D; let $\boldsymbol{S}_D$ denote this sample. Let $\boldsymbol{S}_m$ denote a second sample of size m that could be drawn from the population. An inverse sampling algorithm must describe how to select a sample from $\boldsymbol{S}_D$ so that for any given sample $\boldsymbol{S}_m$

$$\Pr(\text{select } \boldsymbol{S}_m \mid \boldsymbol{S}_D)\ \Pr(\boldsymbol{S}_m \subset \boldsymbol{S}_D) = \Big/ \binom{N}{m}^{1} .$$

The first step is to calculate the probability that an arbitrary but fixed sample $\boldsymbol{S}_m$ is contained in the sample $\boldsymbol{S}_D$. Obviously, there are constraints on the size of the simple random sample (SRS) that can be drawn in this manner; the probability that $\boldsymbol{S}_D$ contains $\boldsymbol{S}_m$ cannot be zero. Certainly, the SRS cannot be larger than the size of the original sample $\boldsymbol{S}_D$, and in fact the size of the SRS is generally required to be much smaller than the original complex sample. The second step is to devise an

algorithm to draw a subsample so that we get the correct probability, Pr(select S_m $|S_D$).

In the 1997 paper, inverse algorithms were provided for a number of common designs. For the problem of interest here, inverting stratified designs and cluster samples are a starting point.

5.2 Inverting A Conventional Stratified Sample. Suppose that we have a stratified sample with fixed sample sizes n_h in each stratum h, and known stratum population sizes, N_h, $\sum N_h$ = N. Because a given sample of arbitrary size m from the population might be contained entirely within one stratum, the largest simple random sample that can be selected from a stratified sample is of size $m = \min\{n_h\}$.

Specifically, assume we have a stratified sample with four strata. To select an SRS of size m from the stratified sample, one must first determine the number of units to be chosen from each stratum. Using a probability distribution generator, select the vector of sample sizes, (m_1, m_2, m_3, m_4), from the hypergeometric distribution where $Pr(m_1=i_1, m_2=i_2, m_3=i_3, m_4=i_4) =$

$$\frac{\binom{N_1}{i_1}\binom{N_2}{i_2}\binom{N_3}{i_3}\binom{N_4}{i_4}}{\binom{N}{m}}$$

where $\sum i_j = m$ and $0 \le i_j \le m$, j=1,...4.

After choosing the pattern of stratum sample sizes, (m_1, m_2, m_3, m_4), select a simple random sample of size m_1 from the n_1 sample units in stratum 1, an SRS of size m_2 from the n_2 sample units in stratum 2, etc.

With some algebra (Hinkins, Oh and Scheuren 1997) it can be shown that this procedure will reproduce a simple random sampling mechanism unconditionally, i.e., when taken over all possible stratified samples. This approach generalizes for any number of strata.

5.3 Inverting a One-Stage Cluster Sample. In this subsection, we consider inverse algorithms for cluster samples where the clusters are sampled by a simple random sampling mechanism and without replacement. We summarize the algorithm for inverting cluster samples where the clusters are of equal size. The more usual case where the clusters are of unequal size is then briefly described; the detailed description for this case can be found in Hinkins, Oh, and Scheuren (1997).

Assume we have a population of N clusters where all clusters are of size M and k of them are selected by a simple random sampling mechanism without replacement. The largest SRS of elements that can be selected is k; the cluster size is not a constraint on the size of the subsample. For a given sample S_k, let q denote the number of clusters represented in S_k; $0<q\le k$. Then the probability that S_k is contained in the cluster sample is equal to the number of cluster samples containing these q clusters divided by the total number of possible cluster samples, i.e.

$$Pr(S_k \subset S_D) = \frac{\binom{N-q}{k-q}}{\binom{N}{k}}.$$

As for the stratified sample, the algorithm first determines the number of units to be chosen from each cluster, (m_1, m_2, ..., m_k). The probability distribution to be used to select the m_i's is $Pr(m_1=i_1,\ldots,m_k=i_k) =$

$$\frac{\binom{M}{i_1}\cdots\binom{M}{i_k}}{\binom{NM}{k}}\frac{\binom{N}{q}}{\binom{k}{q}}$$

where $0\le i_j \le k$, $\sum_j i_j = k$ and q denotes the number of nonzero i_j's. Unlike the stratified example, where the function for selecting the values of m_i was a known probability function, it is not immediately obvious this equation describes a probability distribution. It is shown in the Hinkins, Oh, and Scheuren paper (1997) that this is in fact a probability function.

Once the m_i's are determined, a simple random sample of size m_i is selected from cluster i, i=1,2,...,k. Therefore the conditional probability of selecting S_k is

$$Pr(S_k|S_D) = \frac{1}{\binom{NM}{k}}\frac{\binom{N}{q}}{\binom{k}{q}}.$$

It is then easy to verify that

$$Pr(S_k \mid S_D)\, Pr(S_k \subset S_D) = 1\Big/\binom{NM}{k}$$

and therefore this gives the correct probability of selecting an SRS.

It would appear to be straightforward to generalize this approach in an obvious way to the case of unequal cluster sizes. However, the inverse sampling algorithm for a sample of clusters of equal size does not generalize

readily when a sample of unequal sized clusters is drawn. This difficulty can be fixed, although not perhaps in an entirely satisfactory way. One method is to employ a hypergeometric that assumes all the clusters were as large as the largest cluster in the population. The price paid is that the inverse sample size achieved is no longer fixed, and the resulting subsample is only conditionally SRS given the achieved sample size, denoted, say, as k_0. That is, for a given sample size k_0, $k_0 \leq k$, all samples of size k_0 have the same probability of being selected using the inverse algorithm. A complete discussion can be found in Hinkins, Oh, and Scheuren (1997).

5.4 Multistage Cluster Designs. The design to be inverted is a multistage design. In many cases a multistage design can be inverted by using results from the single stage designs. For example, in the case of a multistage design with PPS sampling at the first stage and SRS sampling at the second, one way to construct an inverse would be to take a **srswr** sample of k clusters and then within each selected cluster take one observation at random. Other inverse algorithms may exist too. A systematic inverse seems reasonable, provided the probability of selecting the same cluster more than once is small to vanishing. In a similar manner, an algorithm for inverting the NHIS design will be constructed.

What about the problem of having only two Primary Sampling Units? From previous discussion, it is immediate that if an inverse is to exist, then the sample size m cannot be any larger than m =2. Depending on the sampling within each stratum, we could employ one or more of the exact or approximate inverses to obtain two SRS selections within each stratum. The inverse algorithm would result in just two selections overall.

How can a method that selects only a sample of size two be of any practical value in the NHIS case? One answer is repeatedly. The next section (Section 6) discusses this briefly, sketching the approach we plan to take.

6. NHIS Application

The NHIS is based upon a highly stratified multistage probability sample. But in order to estimate the variance of the estimators, a simplified design structure is assumed. For the purposes of inverting the sample, we will make the same simplifying assumptions that are used for the variance estimation. The variance estimates using the inverse samples will be compared to the reported variance estimates under these assumptions. The following is a description of this conceptual design based primarily on Chapter 3 of a draft of the NHIS Design Report (National Center for Health Statistics 2000).

6.1 PSUs and Stratification. The survey is stratified at the state level, and the analysis we plan to make will also be at the state level. Primary Sampling Units (PSUs) are defined in a given state as territorial divisions, such as counties or metropolitan areas. The PSUs within the state are stratified where strata are defined using MSA classification and poverty status. There are two types of strata defined: self-representing (SR) strata and non-self-representing (NSR) strata. The largest metropolitan areas are classified as SR strata; in SR strata all PSUs are included in the sample. For the NSR strata, 2 PSUs are generally selected without replacement with probability proportional to population size. (In the smaller states some strata have only one PSU selected. Provisionally, in our planning so far, we expect to employ a collapsed stratum technique before inverting the NHIS design.)

6.2 Substrata and Secondary Sampling Units within PSUs. Each PSU is subdivided into density substrata 1 to 21. Substrata 1 to 20 are defined by joint black and Hispanic concentration measures in block units defined by the 1990 Census. Substratum 21 contains new (post-Census) construction, defined by a continuously updated building permit frame. Most PSUs will not contain blocks in all possible substrata; in fact, most PSUs have only a few such substrata.

Within each substratum, secondary sampling units (SSUs) are defined as clusters of residential housing units. The complexities of the within-PSU sampling require us to make some simplifying sampling assumptions about SSU selection. The SSU will be considered as a well defined population cluster and the SSU sampling treated as having been drawn with replacement sampling from a finite population of SSUs within a given substratum. All SSUs within a substratum will be treated as having the same selection probability, independently selected over substrata, with weights applied to units selected from the SSUs that are assumed to produce an unbiased estimator of the SSU total.

For an SSU selected in the sample, there is then a sampling procedure to select housing units within SSU, collections of individuals within housing units, and finally to select persons within a collection of individuals within a housing unit. But these further details are not described here because they are not used in variance calculations and therefore will not be considered for creating an inverse sample. The unit of analysis will be the estimate for an SSU, which is based on the probabilities of selecting units within that SSU.

6.3 Conceptual Design. Recall that for the SR strata in the NHIS, there is no sample selection of the PSUs; all PSUs are included in the sample. Let i = 1,...N_s denote the PSUs in SR stratum s; let j=1,...21 denote the substratum. We condition on N_{sij}, the number of SSU in SR stratum s, PSU i, substratum j, and the number of SSUs selected, n_{sij}.

For the purposes of our study, the unit of analysis is $y_{sijk} = \hat{Y}_{sijk} = \sum_u x_{sijku} / \pi_{uk.sij}$ where $\pi_{uk.sij}$ is the conditional probability of selecting unit u in SSU k, given that SSU k in stratum s, PSU i, substratum j, has been selected. For variance estimation, y_{sijk} is treated as the population value for SSU k (in substratum j, PSU i, stratum s). In this case, the estimate of the total for stratum s is

$$\hat{Y}_s = \sum_i \sum_j^{c_{si}} \frac{N_{sij}}{n_{sij}} \sum_k^{n_{sij}} y_{sijk} = \sum_i \sum_j N_{sij} \bar{y}_{sij}$$

where c_{si} denotes the number of substratum within PSU i in stratum s. The estimated variance is

$$Var(\hat{Y}_s) = \sum \sum \frac{N_{sij}^2}{n_{sij}} s_{sij}^2$$

where $s_{sij}^2 = \frac{1}{(n_{sij}-1)} \sum_k (y_{sijk} - \bar{y}_{sij})^2$. (Recall that it is assumed that the sampling of the SSUs is **with** replacement.)

In summary, for a self-representing stratum, the design to be inverted is basically a design that selects n_{ij} clusters from a total of N_{ij} clusters from each PSU/substratum ij. The selection of the n_{ij} clusters is with replacement with all clusters in the PSU/substratum having the same probability of selection.

For a non-self-representing stratum the structure for the sample selection within a PSU/substratum is the same for the NSR. However, in a NSR stratum this selection is not performed in every PSU but rather to PSUs selected with sampling proportional to size.

7. Next Steps

Two states have been chosen for our pilot study – a large state and a fairly small one. We are now programming the inverse algorithm and conforming the NHIS sample cases for the two states selected to do the inversions. We plan to do the inverses multiple times since each SRS sample is so small (m=2).

The theory for pooling these sample estimates and calculating the needed variances is all available from Hinkins, Oh and Scheuren (1997), albeit it can be predicted that we will need to develop more theory once we get into the details. For example, we will need to decide how to treat the nonresponse and missing data problems that arise.

We are expecting to be able to demonstrate that the variance stability of our approach will be superior to that using balanced half-sample replicates. We also expect to alleviate the existing concerns about releasing files that can be analyzed by state will be possible -- without compromising respondent confidentiality. Wish us luck; we will need it.

References

Efron, B. (1979). Bootstrap methods: another look at the jackknife, *Annals of Statistics*, 7, 139-172.

Hinkins, S., Oh, H. L., and Scheuren, F. (1997). Inverse Sampling Design Algorithms. *Survey Methodology,* Vol. 23, No. 1, 11-21.

Kennickell, A. (1999). Complete imputation for confidentiality protection of the Survey of Consumer Finances, *Journal of Official Statistics.*

McCarthy, P. and Snowden, C.(1985). The bootstrap and finite population sampling, *Vital and Health Statistics*, Series 2, No. 95, DHHS Pub. No. (PHS) 85-1369.

Mulrow. J. and Scheuren, F. (1999). The Confidentiality Beasties, *Turning Administrative Systems into Information Systems*, Internal Revenue Service.

National Center for Health Statistics, (2000). National Health Interview Survey: design and estimation for the National Health Interview Survey, 1995-2004, *Vital and Health Statistics Series 2*, to appear.

Rubin, D. (1992). Synthetic microdata to assure confidentiality protection, *Journal of Official Statistics.*

Russell, B., Leonard, M. and Scheuren, F. (1999). *Focal Child Public Use File Codebook,* Methodology Report No. 2, 1997 National Survey of America's Families.

Remaining references upon request.