Increasing Public Accessibility to Complex Survey Data by Using Inverse Sampling

Susan Hinkins, Ernst and Young, Van Parsons, National Center for Health Statistics,
and Fritz Scheuren, The Urban Institute
Susan Hinkins, 1122 South 5th Avenue, Bozeman, MT 59715
email: susan.hinkins@ey.com

Keywords: Disclosure avoidance, variance estimation efficiencies

Abstract: In many surveys, geographical location is an important part of the sampling structure, but confidentiality concerns may prohibit the release of the finer details of the geographical sampling structures. Without this information, analysis of the data using the standard design-based methods is very difficult. An inverse sampling algorithm is a mechanism to subsample the original sample data in order to generate a "new" sample that can be treated as a simple random sample from the population. (Hinkins, Oh, and Scheuren, Survey Methodology, 1997.) In this new sample, the geographical identifiers would not be needed for analysis. If inference based on this technique can be demonstrated to be consistent with full complex sample techniques, then data users can perform select design-based analyses using mainstream statistical software. This paper describes how the inverse sample technique could be applied to a typical NCHS type survey.

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 the standard design-based methods.

1.1 Alternatives. An inverse sample algorithm (Hinkins, Oh, and Scheuren, 1997) is a subsampling mechanism on the original sample data that generates a new sample from the original complex sample. In its original application, the resamples drawn were such that they could be treated as simple random selections from the entire population. The goal of the approach is to resample a complex sample to obtain a data structure that is easier to analyze. Because any given resample is unlikely to contain all the information in the original survey, the original complex sample is repeatedly resampled. This paper describes several possible approaches for using the inverse sampling algorithm to enhance a public use file, including using resamples that may have structures other than being simple random in design.

Inverse sampling techniques may provide a useful means to allow the public to have access to micro-level NCHS data because the geographical identifiers would not be needed for the analysis. The inverse sample may also permit implementation of some commonly available data analysis procedures using traditional computer software. If inference based on inverse-sample techniques can be demonstrated to be consistent with full complex-sample techniques, then data users with limited computer resources can perform select design-based analyses using mainstream statistical software.

A public use file is released for the NHIS data in a form where the complex sample structure is simplified to that of a stratified design with two Primary Sampling Units (PSUs) imbedded within each stratum. The original design strata and PSUs were masked in part using some of the techniques discussed in Eltinge (1999) and Parsons and Eltinge (1999). This masked "2 PSUs per stratum" design can be used to calculate variances. For analytical domains covering most of the strata the variance estimators will be stable, i.e., have a large associated degrees of freedom, but for less geographically dispersed domains, covering few strata, the resulting degrees of freedom may be very small, and the variance estimate may be quite unstable.

By drawing many, many samples from the public use design, it may be possible to produce a more stable variance estimator. For less geographically dispersed subpopulations, the inverse sample algorithm could be used as a black-box calculation to provide the user with a more stable variance estimator than the publicly released structure.

1.2 Organization. In the current section (Section 1), we have described the problem and sketched broadly the alternatives we will look at. Section 2 describes the conceptual design for the NHIS and the issues involved with inverting this design. Section 3 describes the NHIS public use file and the final section (Section 4) describes our current plans for using the inverse sampling methodology to improve variance estimates as compared to the method now available on NHIS public use files.

2. Conceptual Design

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 similar simplifying assumptions. We also assume that there are no nonsampling errors such as nonresponse or missing data, which would require special variance treatment. To do this, we will rely on the following description of the conceptual design that has been summarized from information found in Botman, Moore, Moriarity, and Parsons (2000).

2.1 Primary Sampling Unit (PSU) Selection. The survey is stratified at the state level, and the analysis considered in this paper is at the state level. The PSUs are defined as territorial divisions, such as contiguous counties or metropolitan areas and are stratified 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 with certainty. For the NSR strata, 2 PSUs are selected without replacement with probability proportional to population size. (Actually, for most of the NSR strata two PSUs are selected, but for some smaller states only one PSU is selected. However, in this paper we confine attention to the states where 2 PSUs are selected in all NSR strata.)

2.2 Choosing Secondary Sampling Units (SSUs). Each PSU is subdivided into 21 density substrata. Substrata 1-20 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.

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. Each substratum's universe is considered to be a set of well-defined population clusters, the SSUs. The SSU sampling is treated as sampling from a finite population of known SSUs within a substratum. All SSUs within a substratum have the same selection probability and sampling is independent over substrata. It is also assumed that the weights applied to units selected from within the SSUs produce an unbiased estimator of the SSU total.

2.3 Choosing Housing Units (HUs) For an SSU selected in the sample, all black and Hispanic households are retained in the sample, while the complement is subsampled. There is then a sampling procedure to select individuals within selected housing units, and finally to select persons within a collection of individuals within a housing unit. These finer details of the sample design are actually not used in variance calculations; the unit of analysis for variance estimation is based on the estimate for an SSU aggregation, which is based on the probabilities of selecting units within that SSU. For first-order estimation, the fine household or individual level design information will be used to establish the weights.

3. Public Use Data Files.

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 the sacred oaths made to respondents to preserve the confidentiality of any data entrusted to the agency. 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. The challenge is to preserve the analytic potential of the publicly available data, while preserving the confidentiality promised to respondents.

3.1 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. 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. Those 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.

3.2 Original Resampling Approach. The original purpose of this project was to investigate the possibility of using the inverse sampling algorithm to provide data that could be made publicly available and that could be analyzed without knowing the geographically linked information. We believe that it is theoretically possible to invert the NHIS design, as described in Section 2, to the elementary sampling unit. However, in this case, the largest simple random sample that could be drawn, even if drawing only down to the household level, would be a sample of size 1. Even with many, many resamples, the estimation of variance from resamples with m=1 is not feasible.

A possible revised goal would be to invert the conceptual design at the level that the variances are now calculated (where the SSU is the unit of analysis). This would mean drawing many resamples of SSUs. But in terms of providing a public use file that protected confidentiality, this would result in micro-data only to the SSU level- data aggregated to clusters of households. And this approach was not judged feasible either, unless modified in some way. We return to this below.

3.3 Pseudo PSUs. Many surveys now implement masking techniques to define the strata and PSUs imbedded within a public use file. These structures are constructed in such a way as to produce variance estimators that have approximately the same expectations as those produced by a full design information variance estimator, but at the expense of a loss of degrees of freedom.

Taking one geographic region as an example, the public use file contains 27 strata with two PSUs per stratum. For sub-populations of interest that are found in all strata, the variance estimate would have 27 degrees of freedom and would be fairly stable. However, for relatively rare characteristics that are not spread out uniformly over the PSUs, the variance estimate may have very few degrees of freedom. For example, for certain classifications of race/ethnicity, there are only two strata containing individuals in both replicates, and eight additional strata with an individual in just one of the two replicates. For cases such as this, we propose that by using the inverse sampling algorithm to draw multiple simple random samples from the underlying design, a more stable variance estimate can be calculated.

3.4 Conceptual Design to Invert As described earlier, the sampling unit for inverting the design is the Secondary Sampling Unit which is a cluster of households. For strata that correspond to the original NSR strata, the design to be inverted is a multistage design. We again have to make some simplifying assumptions in order to invert. At the first stage, in each stratum, 2 PSUs are selected using ppswr. Within each sampled PSU, a stratified design is used to select a simple random sample of n_h SSUs is selected from the N_h population SSUs.

Using results from Hinkins, Oh, and Scheuren (1997) this design can be inverted as follows. Within each stratum, s, 2 PSUs are selected; therefore the largest simple random sample that can be selected is of size 2. The design to be inverted is a multi-stage sample: a stratified design with K_s PSUs in stratum s, where PSU i contains M_{si} SSUs. At the first stage, 2 PSUs in stratum s are selected using ppswr sampling, where size is measured by M_{si}. In each selected PSU i, sub-strata are defined and a srs of size m_{sij} is drawn in substratum j, PSU i, stratum s, from the known M_{sij} SSUs. Conceptually, the design can be inverted by starting at the last stage and inverting each piece. The stratified design within each PSU can be inverted to produce a simple random sample of size, $n_{si}=\min(m_{sij})$. Then within each stratum, we have a multi-stage design with pps sampling at the first stage and srs sampling at the second. This can be inverted by selecting a srswr sample of k=2 PSUs and then within each selected PSU, take one observation at random. Now we have a stratified sample where within each stratum, we have a srs of size 2 out of K_s. This can be inverted as described in the 1997 paper.

We believe that the strata constructed from the original SR strata can be treated similarly, except at the first stage, the stratum contains two PSU's, both of which are selected in the sample. This removes one step from the algorithm for inverting these strata. For simplicity, of course, we could continue to treat the SR portion of the sample, just like the NSR portion, since with enough resamples, the two approaches will still be equivalent.

4. Inverse Sampling

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. By repeating the entire subsampling procedure, we can generate g simple random samples each of size m, where each SRS is selected independently from the overall original sample. Each repetition must include all steps of the subsampling procedure.

These inverse sampling algorithms, when feasible, make it possible to employ conventional techniques, like regression and contingency table analysis, with only minor adjustments. Furthermore, in the 1997 paper, conditions are given under which the precision of the estimates using multiple SRS's can be made arbitrarily close to the precision of the original estimates. For many estimators, these conditions can be met or the conditions can be met for Taylor series approximations to the variance.

Therefore, for estimation of rare events from the public use file, it should be possible to calculate more stable variance estimates by resampling many, many simple random samples via the inverse sampling algorithm. One could conceivable provide the user with the many resamples at the SSU level only, or more likely one would build a black box using inverse sampling techniques to provide the user with an improved variance estimate. The future work will be to draw many simple random samples of SSUs, using the inverse algorithm and evaluating the improvement in stability in the variance estimate for rare sub-populations.

Research results and conclusions expressed are those of the authors and do not necessarily indicate concurrence by the National Center for Health Statistics. This paper is released to encourage discussion.

References

Botman, S.L., Moore T.F., Moriarity, C.L., and Parsons, V.L., Design and estimation for the National Health Interview Survey, 1995-2004. ***National Center for Health Statistics. Vital Health Stat*** *2*(130), 2000.

Eltinge, J. L. (1999) Use of stratum mixing to reduce primary-unit-level. identification risk in public-use survey datasets. **Proceedings of the 1999 Federal Committee on Statistical Methodology Research Conference**

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

Parsons, V.L. and Eltinge, J.L. (1999). Stratum partition, collapse and mixing in construction of balanced repeated replication variance estimators. **American Statistical Association, Proceedings of the Section on Survey Research Methods**, pp. 843--848.