Survey Methodology, June 1997
Vol. 23, No. 1, pp. 11-21
Statistics Canada

Inverse Sampling Design Algorithms

SUSAN HINKINS, H. LOCK OH and FRITZ SCHEUREN[1]

ABSTRACT

In the main body of statistics, sampling is often disposed of by assuming a sampling process that selects random variables such that they are independent and identically distributed (IID). Important techniques, like regression and contingency table analysis, were developed largely in the IID world; hence, adjustments are needed to use them in complex survey settings. Rather than adjust the analysis, however, what is new in the present formulation is to draw a second sample from the original sample. In this second sample, the first set of selections are inverted, so as to yield at the end a simple random sample. Of course, to employ this two-step process to draw a single simple random sample from the usually much larger complex survey would be inefficient, so multiple simple random samples are drawn and a way to base inferences on them developed. Not all original samples can be inverted; but many practical special cases are discussed which cover a wide range of practices.

KEY WORDS: Finite population sampling; Inference in complex surveys; Resampling.

1. INTRODUCTION

The development of modern survey sampling is an extraordinary achievement (Bellhouse 1988; Hansen 1987; Kish 1995). The very richness in that development may have had the effect, though, of isolating survey sampling from the rest of statistics – where it is the richness of models that is given emphasis. In fact, it is a well-known commonplace that, in the main body of statistics, sampling is often disposed of by assuming a sampling process that selects random variables such that they are independent and identically distributed (IID).

Important techniques, like regression and contingency table analysis, were developed largely in this IID world; hence, adjustments are needed to use them in complex survey settings. Indeed, whole books have been written on this problem (Skinner, Holt and Smith 1989); and much time and effort have been devoted to it in software (like SUDAAN or WESVAR PC) specially written for surveys (See also Wolter 1985). With all that has been done already, can something more of value be added? We think we may have a contribution to offer on how to deal better with the "seam" which currently exists between IID and survey statistics.

Organizationally, the paper is divided into four sections. This introduction is Section 1. In Section 2 and 3 a general problem statement is provided and several "resolutions" are offered in a few of the better known designs. Our approach is to resample the complex sample to obtain an easier to analyze data structure. Specifically, we cover stratified element sampling, one and two-stage cluster samples, plus the important two PSU per stratum design (Section 2). Because any given resample is unlikely to contain all the information in the original survey, we look at what happens when the original complex sample is repeatedly resampled. A concrete illustration of our ideas is also given in Section 3; this has been taken from our practice and is based on a highly stratified Statistics of Income (SOI) sample of corporate tax returns (*e.g.*, Hughes, Mulrow, Hinkins, Collins and Uberall 1994). In a concluding section (Section 4), we discuss a few applications and some next steps needed for our still embryonic ideas to grow more useful.

2. PROBLEM STATEMENT AND POSSIBLE "RESOLUTIONS"

2.1 Motivation and Basic Approach

Suppose we wanted to apply an IID procedure to a complex survey sample. Suppose, too, that we wanted to take a fresh look at "solving" the seam problem that occurs because the survey design is not IID. How might one proceed? Well, there is a familiar expression that may fit our approach

**If you only have a hammer, every
problem turns into a nail.**

Now, as samplers, we have a hammer and it is sampling itself. Can we turn the seam problem in surveys into a nail that can be dealt with by using another sampling design?

It is our contention that some of the time the answer to this question is "Yes." We call this second sample design an "Inverse Sampling Design Algorithm" – hence, the name of this paper.

A schematic might help visualize the algorithm (see figure 1). In the diagram two sampling approaches are compared – both yielding simple random samples from a population:

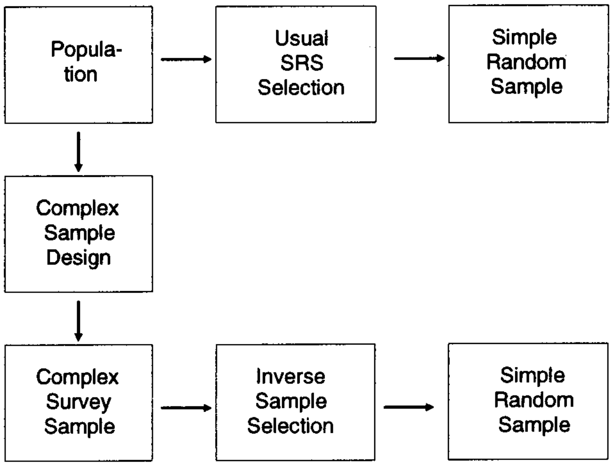
(1) The first design (top row) does this by employing a conventional direct simple random (SRS) selection process (*e.g.*, Cochrane 1977), such that all possible

[1] Susan Hinkins, Internal Revenue Service, Bozman, MT, U.S.A.; H. Lock Oh, Internal Revenue Service, Washington, DC, U.S.A.; Fritz Scheuren, Ernest and Young, 1402 Ruffner Rd., Alexandria, VA 22302 U.S.A.

samples of a given size have the same probability of selection. (Such designs are often impracticable or inefficient or both; hence, they are almost never used by survey samplers, despite their ubiquity in textbooks.)

(2) The second design envisions a two-step process. The first step is to sample the population in a complex way that focuses carefully on the nature of the population and the client's needs – using the client's resources frugally (this is the survey sampler's province, par excellence). '

(3) What is new in our formulation is to draw a second (perhaps complex?) sample that inverts the first set of selections, so as to yield at the end a simple random sample. Of course, to employ this two-step process to draw a single simple random sample from the usually much larger complex survey would be inefficient, so we propose to create multiple simple random samples and base our inferences on them.



While elaborations are possible, the basic nature of the algorithms we are talking about should, by this point, be obvious. They can consist of just four basic steps:

(1) Invert, if you can, the existing complex design, so that simple random subsamples can be generated (to some useful degree of approximation).

(2) Potentially, apply your conventional statistical package directly to the subsample, since that is now appropriate.

(3) Repeat the subsampling and conventional analysis, in steps (1) and (2), over and over again.

(4) Retain, if you can, the flavour of the original randomization paradigm by using the distribution of subsample results as a basis of inference (rather than the original complex sample).

Notice some things that this approach is – and is not: First, it is extremely computer intensive – presupposing cheap, even very cheap computing. 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 Snowmen 1985; Rao and Wu 1988) – just the opposite; our goal here is to create a totally different and more analytically tractable set of subsamples from the original design.

2.2 Defining An Inverse Sampling Algorithm

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 S_D denote this sample. Let 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 S_D so that for any given sample S_m

$$\Pr(\text{select } S_m \mid S_D) * \Pr(S_m \subset S_D) = \frac{1}{\binom{N}{m}}. \tag{1}$$

The first step is to calculate the probability that an arbitrary but fixed sample S_m is contained in the sample 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 S_D contains S_m cannot be zero. Certainly, therefore, the SRS cannot be larger than the size of the original sample S_D, and in fact the size of the SRS is generally required to be much smaller than the original complex sample.

The problem, then, is to find a general algorithm to select an SRS from a given sample S_D with the correct conditional probability. It is also necessary to check that valid probability functions are used. The following subsections show the inverse sampling algorithms for a few of the more common sample designs: stratified, cluster, multistage, and stratified multistage designs. We also give an example where an inverse algorithm at first does not appear feasible.

2.3 Inverting A Stratified Sample

In this subsection the inverse algorithm is given for a stratified sample with four strata. The algorithm generalizes for any number of strata. We have a stratified sample with fixed sample sizes n_h in each stratum h, and known stratum population sizes, $N_1 + N_2 + N_3 + N_4 = N$. 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 $m = \min\{n_h\}$.

For a given sample S_m, let (x_1, x_2, x_3, x_4) denote the number of units in each stratum. Each x_i will be between 0 and m, and $x_1 + x_2 + x_3 + x_4 = m$. The probability that S_m is contained in the stratified sample is equal to the number of stratified samples containing these m units divided by the total number of possible stratified samples, *i.e.*

$$\Pr(S_m \subset S_D) = \frac{\binom{N_1-x_1}{n_1-x_1}\binom{N_2-x_2}{n_2-x_2}\binom{N_3-x_3}{n_3-x_3}\binom{N_4-x_4}{n_4-x_4}}{\binom{N_1}{n_1}\binom{N_2}{n_2}\binom{N_3}{n_3}\binom{N_4}{n_4}}. \qquad (2)$$

The algorithm for selecting a SRS from the stratified sample consists of the following three steps:

(1) Determine the size of the SRS to be selected: $m \le \min\{n_h\}$.

(2) Generate a realization $\{m_1, ..., m_4\}$ from a hypergeometric distribution, with probabilities

$$\Pr(m_1 = i_1, m_2 = i_2, ..., m_4 = i_4) = \frac{\binom{N_1}{i_1}\binom{N_2}{i_2}\binom{N_3}{i_3}\binom{N_4}{i_4}}{\binom{N}{m}} \qquad (3)$$

where $i_1 + i_2 + i_3 + i_4 = m$ and $0 \le i_1 \le m, 0 \le i_2 \le m$, $0 \le i_3 \le m$, $0 \le i_4 \le m$.

(3) In each stratum h, select a simple random sample of size m_h, without replacement, from the n_h sample units.

The conditional probability of selecting the sample S_m given that it is contained in the stratified sample, is then

$$\frac{\binom{N_1}{x_1}\cdots\binom{N_4}{x_4}}{\binom{N}{m}} \frac{1}{\binom{n_1}{x_1}\cdots\binom{n_4}{x_4}}. \qquad (4)$$

The probability of selecting any given sample S_m using the inverse algorithm is the product of the two probabilities given in equations (2) and (4). It is straightforward to show that this product is equal to

$$\frac{1}{\binom{N}{m}}.$$

Therefore this procedure reproduces a simple random sampling mechanism unconditionally, *i.e.*, when taken over all possible stratified samples. Note that in order to generate all possible SRS's from this population, the entire sequence must be repeated, starting with selecting a stratified sample and proceeding through steps 1 - 3.

2.4 Inverting a One Stage Cluster Sample

In this subsection, we consider three special cases. To begin with, we examine cluster samples where the clusters are of equal size. This is followed by the more usual case where the clusters are of unequal size. In both of these settings we assume the clusters are sampled by a simple random sampling mechanism and without replacement. The third case studied is that of sampling unequal clusters by a probability proportional to size (PPS) mechanism. In this last instance we assume that the sampling is with replacement.

2.4.1 One Stage Cluster Sampling With Equal Cluster Sizes, Sampled With Equal Probability

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.

To construct an inverse algorithm, we need to decide what the largest element subsample might be. It is immediate that the largest SRS of elements that can be selected is k. Incidentally, 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.*

$$\Pr(S_k \subset S_D) = \frac{\binom{N-q}{k-q}}{\binom{N}{k}}. \qquad (5)$$

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

$$\Pr(m_1=i_1,...,m_k=i_k) = \frac{\binom{M}{i_1}\cdots\binom{M}{i_k}}{\binom{NM}{k}} * \frac{N(N-1)...(N-q+1)}{k(k-1)...(k-q+1)} \qquad (6)$$

where $0 \le i_j \le k$, $i_1 + i_2 + ... + i_k = k$, and q is the number of nonzero i_j's. For example, with $M = 100$, $N = 6$, and $k = 3$

$$\Pr(m_1 = 1, m_2 = 0, m_3 = 2) = \frac{\binom{100}{1}\binom{100}{0}\binom{100}{2}}{\binom{600}{3}} * \frac{6*5}{3*2}$$

$$\Pr(m_1 = 3, m_2 = 0, m_3 = 0) = \frac{\binom{100}{3}}{\binom{600}{3}} * \frac{6}{3}.$$

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

$$\Pr(\text{select } S_k \mid S_D) = \frac{1}{\binom{NM}{k}} * \frac{N(N-1)...(N-q+1)}{k(k-1)...(k-q+1)}. \quad (7)$$

The probability of selecting a particular sample S_k is found by multiplying equation (5) times equation (7). It is routine to verify that this gives the correct probability of selecting an SRS.

Unlike the stratified example, where the function for selecting the values of m_i was a known probability function, it is not immediately obvious that equation (6) describes a probability distribution. Since the values generated by this function are all nonnegative, it need only be shown that they sum to one over the space of possible values. The first factor in the equation has the form of a hypergeometric distribution, except that the numerator is constrained to only k out of the N clusters, while the denominator still reflects the total N clusters. It is useful to define a partition of k as a combination of positive integers that adds to k, without regard to order. For example, the partitions of $k = 3$ are $\{3\}$, $\{1,2\}$, and $\{1,1,1\}$. Because the clusters are all of the same size, M, all patterns of selection that correspond to the same partition have the same probability of occurring. Take, for example, $N = 6$, and $k = 3$. In the full hypergeometric distribution, with equal cluster size, each of the following combinations has the same probability of occurring

$$(0,0,0,0,1,2), (0,0,0,0,2,1), (0,0,0,1,2,0), ..., (2,1,0,0,0,0).$$

The total number of such combinations is $N(N-1) ... (N-q+1)$, where q is the size of the partition, that is the number of (nonzero) values in the partition. In the example above, $q = 2$. For a given partition, if the nonzero counts can only be put into k specific cells, then there are $k(k-1) ... (k-q+1)$ such orderings. Therefore, summing the distribution over all values of $(i_1, ..., i_k)$ can be done by first summing over all partitions of k and then for each partition, summing over all possible orderings of that partition in k cells. Because all orderings associated with a particular partition share a common probability of occurrence, this results in a summation that is equivalent to summing the hypergeometric over the correct space, and therefore expression (6) sums to one.

The probability distribution needed for this simple cluster design (equation 6) is noticeably more difficult to generate than the hypergeometric distribution in the case of the stratified sample. However, as the sampling fraction k/N decreases, the probability is often contained in only two of the partitions: $q = k$ and $q = k - 1$. (These probabilities are calculated in the Appendix). Indeed, the probability may be concentrated in just the pattern with $q = k$ (A special case of this is also shown in the Appendix).

Given the results in the Appendix, it may be possible to approximate the exact inverse by selecting one case from each cluster, using systematic sampling from the original cluster sample. This approach is of real value because the probability distribution calculations become unwieldy as the number of clusters in the sample grows large. For a systematic inverse to work, however, the "step" would naturally have to be at least as large as M or maybe even greater, depending on the number of clusters in the population. To carry out this subsampling repeatedly, for each systematic sample inverse, the units within each cluster would be reordered randomly before the next selection and the clusters resorted randomly as well - then another random start obtained before stepping again through the original sample.

2.4.2 One Stage Cluster Sampling with Unequal Clusters, Sampled With Equal Probability

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 is so despite the fact that it would appear to be straightforward to generalize this approach in an obvious way. In particular, it does not seem difficult to generalize the previous method so that the "probabilities" would multiply out successfully to give the "correct" probability of selection, *i.e.*

$$\frac{1}{\binom{M_+}{k}}, \quad \text{where } M_+ = \sum_1^N M_i. \quad (8)$$

However, generalizing to unequal cluster sizes M_i by selecting the m_i as

$$\Pr(m_1{=}i_1,...,m_k{=}i_k) = \frac{\binom{M_1}{i_1}...\binom{M_k}{i_k}}{\binom{\sum_1^N M_i}{k}} * \frac{N(N-1)...(N-q+1)}{k(k-1)...(k-q+1)} \quad (9)$$

does not result in a valid probability distribution. We will again assume, by the way, that the original clusters are being sampled with equal probability and without replacement, as was the case in subsection 2.4.1. Later (Subsection 2.4.3), as already noted, we will look at original samples which employ some form of Probability Proportional to Size (PPS) selection.

To see that it is not straightforward to simply generalize equation (6) into the form in equation (9), consider the following counter-example where the "probability" calculated using (9) is greater than one. Suppose $N = 4$ with cluster sizes; $M_1 = 4, M_2 = 6, M_3 = 8$, and $M_4 = 10$. Suppose further that we draw a cluster sample with $k = 2$ and that just by chance the two clusters picked are the largest – *i.e.*, $M_3 = 8$ and $M_4 = 10$. It is immediate that with these selections, equation (9) would generate a probability of selecting one unit from each cluster that was greater than one.

Can this difficulty be fixed? Yes, 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, $k_0 \le k$, all samples of size k_0 have the same probability of being selected using the inverse algorithm.

Let M_* denote the maximum cluster size, $M_* = \operatorname{Max}\{M_1, M_2, ..., M_N\}$. Create a population by filling out each original cluster with "dummy" units or placeholders, $j = M_i + 1, M_i + 2, ..., M_*$. Then using a method similar to Lahiri's (1951) for PPS sampling, the inverse algorithm selects units from the population consisting of N clusters each of size M_*, and then discards any element not in the "subpopulation" consisting of the original clusters of size M_i.

Specifically, given a cluster sample consisting of k clusters, select the vector $\boldsymbol{m}$ from the probability distribution

$$\Pr(m_1 = i_1, ..., m_k = i_k) =$$

$$\frac{\binom{M_*}{i_1}\binom{M_*}{i_2}\cdots\binom{M_*}{i_k}}{\binom{NM_*}{k}} * \frac{N(N-1)...(N-q+1)}{k(k-1)...(k-q+1)} \quad (10)$$

where the components of $\boldsymbol{m}$ sum to k, and q of the components m_i are nonzero. This is now a proper probability distribution. Given the selected value of m_i, select a random sample of m_i units from cluster i, where the cluster contains M_i units from the population and $M_* - M_i$ "placeholders." Discard any selected units that are placeholders, in the set of $j = M_i + 1, M_i + 2, ..., M_*$. Therefore the final sample size will not necessarily be equal to k, but may be smaller, say k_0.

The resulting sample is conditionally a SRS from the population, in the sense that for a given value of k_0, all samples of size k_0 have the same probability of being selected using this inverse algorithm. To see this, continue to view the problem as a subpopulation, P, of N clusters of size M_i, $i = 1, ..., N$, within a population P_* of N clusters each of size M_*. Note that for any sample, S_*, of size k selected from the population P_*, the probability of selecting S_* using the inverse algorithm is

$$\frac{1}{\binom{NM_*}{k}}. \quad (11)$$

If $k_0 = k$ then this is the probability of selecting this sample using the inverse algorithm. For a fixed $k_0 < k$, let S_0 denote any given sample of size k_0 contained in P. We can generate a sample S_* containing S_0 by starting with S_0 and adding to it $k - k_0$ elements from the $N * M_* - M_+$ placeholders in P_*. The number of such samples S_*, that result in selecting S_0, is

$$\binom{NM_* - M_+}{k - k_0} \quad \text{where} \quad M_+ = \sum_{i=1}^{N} M_i. \quad (12)$$

Therefore, the probability of selecting S_0 using the inverse algorithm is equal to the probability of selecting S_* using the inverse algorithm, given in (11), summed over all samples S_* constructed as described above, where the number of such samples is given by (12). This probability equals

$$\frac{\binom{NM_* - M_+}{k - k_0}}{\binom{NM_*}{k}}$$

and all samples of size k_0 have the same probability of being selected using the inverse algorithm.

There is a positive probability, unfortunately, that a sample might be selected with this approach that has no elements. This could occur if there were a large difference in the cluster sizes. However, if the number of clusters k in the original sample is large, this is unlikely to be a problem.

Again, as in the case of equal cluster sizes, an approximation is available using a systematic subsample as an inverse. This time we would want a step at least as large as the maximum cluster size. Using a systematic inverse, by the way, would have the advantage of controlling better the actual subsample size drawn.

2.4.3 One Stage Cluster Sampling with Unequal Clusters, Sampled With Unequal Probability

If a sample of k clusters is selected with PPS, an inverse algorithm may exist. Suppose the samples are selected with replacement from a population consisting of N clusters, with unequal cluster sizes, $M_1, M_2, ..., M_N$. Suppose, further, that the measure of size is either equal to M_i or proportional to M_i. Then at each draw,

$$\Pr(\text{select cluster } j) = \frac{M_j}{M_+}$$
$$\text{where } M_+ = \sum_{i=1}^{N} M_i. \quad (13)$$

Finally, since a one stage sample is being taken, once cluster j is selected, then all M_j units from that cluster are included in the sample.

An inverse algorithm in this case should result in a SRSWR. That is, for any vector $\boldsymbol{S}$ resulting from k independent selections from the population, the probability of selecting the ordered vector is

$$\Pr(\text{select } \boldsymbol{S}) = \left(\frac{1}{M_+}\right)^k. \quad (14)$$

An inverse algorithm is to simply randomly select one unit from each cluster in the cluster sample. Because the clusters were chosen with replacement, one should think of the sampled clusters as being ordered, by the order in which they were selected, or in any fixed order. For example, if the population contained 20 clusters, a possible cluster sample of size $k = 5$ is (7, 5, 7, 18, 6), *etc.*

The population consists of M_+ units, denoted as $u_1, u_2, ..., u_{M+}$. Let S denote a given sample, with replacement, $S = (s_1, s_2, .,., s_k)$, and let $\boldsymbol{c} = (c_1, c_2, ..., c_k)$ denote the associated cluster for each unit. For example, suppose the population is:

Cluster	Units
1	$u_1\ u_2\ u_3\ u_4$
2	$u_5\ u_6\ u_7\ u_8$
3	$u_9\ u_{10}\ u_{11}$
4	$u_{12}\ u_{13}\ u_{14}$
5	$u_{15}\ u_{16}\ u_{17}$
6	$u_{18}\ u_{19}\ u_{20}$

and $k = 3$. Then the sample $(s_1 = u_2, s_2 = u_4, s_3 = u_{17})$ corresponds to $\boldsymbol{c} = (1, 1, 5)$. The sample $(s_1 = u_{18}, s_2 = u_{19}, s_3 = u_{18})$ corresponds to $\boldsymbol{c} = (6, 6, 6)$. Note that this second sample can only be selected if cluster 6 is the only cluster chosen in the cluster sample.

For a given sample S of size k, and the corresponding vector $\boldsymbol{c}$ of cluster membership, the unconditional probability of selecting S using the inverse algorithm is

$$\Pr(\text{select } S \mid \text{cluster sample } \boldsymbol{c}) * \Pr(\text{select } \boldsymbol{c}) = \left(\prod_{i=1}^{k} \frac{1}{M_{c(i)}}\right)\left(\prod_{i=1}^{k} \frac{M_{c(i)}}{M_+}\right) \tag{15}$$

which is equal to the desired probability, equation (14).

Note that this same inverse algorithm works in the case where k clusters are selected with ppswr, but a sample of fixed size m is selected (srswor) from the chosen cluster, assuming that $M_i > m$ for all clusters i.

2.4.4 Some Comments On One Stage Designs.

We have seen that, with care, inverse algorithms can be constructed for several special cases where the original sample has a one stage cluster design. Two of our results are for cluster samples drawn with equal probability without replacement. The third is a ppswr design.

A convenient systematic inverse may even be workable as an approximation to the correct inverse algorithm when we have a cluster sample. The approximation works when using SRSWR is "close to" SRSWOR – *i.e.*, in our notation when k/NM is very small so that $1/(NM - k + 1)$ is approximately equal to $1/NM$. So everything seems intuitively to be consistent, across the cases studied.

Many cluster designs do not fall into any of the special cases examined. For some of them we conjecture that exact inverse algorithms may not exist. In particular, the general case of PPSWOR sampling seems to be one of these, including the frequently used variant of systematic PPSWOR. This may, or may not be a problem for practitioners who often employ the (usually) conservative practice of assuming that the sampling was with replacement – in which case an inverse algorithm would exist to the same order of approximation as was being assumed to estimate variances.

2.5 Multistage Cluster Designs

What about multistage designs? Can they be inverted? In some cases, we believe the answer is "Yes." Three designs will be looked at: (1) a two-stage design with simple random sampling at the first and second stages (Subsection 2.5.1); then, (2) a design which employed probability proportional to size (PPS) sampling at the first stage and simple random sampling at the second (Subsection 2.5.2). Finally, (3) the very important stratified multistage design with two PSUs per stratum deserves at least a brief comment.

As will be seen, the stratified and one stage results extend fairly readily. To demonstrate this, our basic strategy is to repeatedly apply the approaches already discussed earlier.

2.5.1 Multistage Designs With Simple Random Sampling at Both Stages

Suppose, first, that originally a simple random sample of k clusters, all of size M, was drawn at the first stage and a simple random subsample of size "r" was drawn at the second stage, within each cluster selected at the first stage.

As earlier, our inverse sample can be no larger than k. Suppose first that $1/(NM - k + 1)$ is approximately equal to $1/NM$, then we can employ an srswr inverse algorithm, since SRSWR and SRSWOR are very close. Using the results in Subsection 2.4.3, we would take a SRSWR sample of k clusters and then within each selected cluster take one observation at random. Alternatively, we could as in Subsection 2.4.1, first determine the number of units to be chosen from each cluster, $(m_1, m_2, ..., m_k)$. Once the m_i's are determined, a simple random sample without replacement of size m_i is selected from cluster i, $i = 1, 2, ..., k$. This may be a nearly exact result, except for the possibility that the inverse second stage sample size m_i may be larger than the original second stage sample size "r." When this occurs, we still can appeal to the results in Subsection 2.4.2 and draw our second stage sample with "placeholders." In this second instance, the resulting actual sample would no longer be fixed; but still would be conditionally SRS. If the first stage clusters are unequal in size but sampled with replacement, then we can again employ the trick used in Subsection 2.4.2 of creating "placeholders." The sample sizes are random and only conditionally do we achieve an SRS inverse.

Another way to approach this problem is to note that the largest SRS that can be selected using an inverse algorithm is of size $k_0 = \min\{k, r\}$. This is done by first determining the number of units to select from each cluster, $(m_1, m_2, ..., m_k)$, where now the m_i's must sum to k_0 rather than k. Once the m_i's are determined, a simple random sample of size m_i is selected from cluster i, $i = 1, 2, ..., k$. The probability distribution to be used to select the m_i's is

$$\Pr(m_1 = i_1, ..., m_k = i_k) = \frac{\binom{M}{i_1} \cdots \binom{M}{i_k}}{\binom{NM}{k_0}} * \frac{N(N-1)...(N-q+1)}{k(k-1)...(k-q+1)}$$

where $0 \le i_j \le k_0$, $i_1 + i_2 + ... + i_k = k_0$, and q is the number of nonzero i_j's.

One final comment, for both equal and unequal cluster sizes, the possibility of an approximate systematic inverse seems available – with essentially the same caveats, of course, as noted above.

2.5.2 Multistage Designs With PPS Sampling at the First Stage and SRS Sampling at the Second

Again, our inverse sample can be no larger than k. It is immediate that one way to construct an inverse would be to use the results in Subsection 2.4.3. Specifically, we would 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.

2.5.3 Stratified Multistage Designs With Two PSU's Per Stratum

Can two Primary Sampling Unit (PSU) designs be inverted? Our answer is "Yes," if the within stratum selections are made in one of the ways we discussed in detail earlier. This is basically the only case we will cover.

From our results in Subsections 2.3 and 2.4, 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 strata, we could employ one or more of the exact or approximate inverses to obtain two SRS selections within each stratum. To obtain an overall SRS sample, we would employ the inverse algorithm of Subsection 2.3 on these two selections and end up, finally, with just two selections overall.

2.5.4 Some Comments On Multistage Designs

In this Subsection, we have quickly covered a few multistage designs and provided exact or approximate inverses. The results were derived by appealing to earlier results in Subsections 2.3 and mainly 2.4. Of course, many multistage designs do not fall into any of the special cases examined - notably those with systematic selections at the last stage.

One last observation, many readers may wonder, at this point, how a method that selects only a sample of size two (as we did in Subsection 2.5.3) can be of any practical value. Perhaps the next section will help.

3. RESAMPLING TO INCREASE POWER

3.1 General Setting

Drawing a single, smaller simple random sample from a larger, more complex sample might be adequate for some users in some settings. However, for most users, the loss in power between the estimate based on the complex sample and the estimate based on a simple random sample would not be acceptable.

In order to increase the power of our approach, it was natural to consider resampling techniques. We are limited in the size of the SRS that can be drawn, but we can repeat the process. 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. For example, in the stratified case, the stratum subsample sizes must be redrawn using the hypergeometric distribution.

In this section, conditions are given under which the precision of the estimates using multiple SRSs can be made arbitrarily close to the precision of the original estimates. We will begin our discussion by first defining some notation.

Let D denote any invertible design (such as a design of the type covered in Section 2). Let T be the population quantity of interest (say, a population total); and let T_D be an unbiased estimator of T calculated from the sample S_D. Suppose g simple random samples are independently drawn from the given sample S_D and let t_i denote the estimator from the i-th simple random sample. Then it can be shown that

$$\text{if } E(t_i \mid S_D) = T_D$$

$$\text{then } \operatorname{Var}\left(\frac{1}{g}\sum_{i=1}^{g} t_i\right) = \operatorname{Var}(T_D) + \frac{1}{g}(\operatorname{Var}(t_1) - \operatorname{Var}(T_D)).$$

Proof: Because the g replications of the simple random sampling process are conditionally independent, then

$$\text{for } i \ne j, E(t_i t_j \mid S_D) = T_D^2.$$

Therefore, unconditionally, for i not equal to j,

$$\begin{aligned}\operatorname{Cov}(t_i, t_j) &= E(t_i t_j) - T^2 \\ &= \operatorname{Var}(T_D).\end{aligned}$$

And the result follows directly.

Some of the conditions in this proof can be relaxed; if T_D is biased, then similar results can be obtained for MSE instead of variance. However, the condition that

$$E(t_i \mid S_D) = T_D$$

is necessary. And this condition is not met for ratio estimators. But, if the condition is met separately for the numerator and for the denominator of the ratio estimate and if the final size of the *combined* sample is sufficiently large so that a Taylor Series approximation is acceptable, then similar results can be found for approximations to the variance for ratios in the usual manner. Incidentally, even in the two PSU per stratum design, this approach works – provided we can obtain an unbiased estimate from each individual sample of size 2. And for estimates of totals, this can be the case – assuming at each stage of sampling that an inverse can be constructed.

3.2 Estimating The Sampling Error for Means or Totals

By resampling, one can achieve almost the same precision as the original design estimator. But because the resampled srs's are only conditionally independent, the estimation of the standard error is not as simple as if only one srs had been drawn. However the estimation remains relatively straightforward.

Let S^2 denote the population variance for the variable X and let T be its population total. For the sample means, totals and variances calculated from the generated simple random samples, let

$$t_{**} = \frac{1}{g}\sum_{j=1}^{g} t_j = \frac{1}{g}\sum_{j=1}^{g} N\bar{x}_j = \frac{1}{g}\sum_{j=1}^{g}\frac{N}{m}\sum_{i=1}^{m} x_{ji}$$

$$s_j^2 = \left(\frac{1}{m-1}\right)\sum_{i=1}^{m}(x_{ji} - \bar{x}_j)^2$$

$$s_*^2 = \left(\frac{1}{gm-1}\right)\sum_{j=1}^{g}\sum_{i=1}^{m}(x_{ji} - \bar{x}_{**})^2$$

where $\bar{x}_{**} = \frac{t_{**}}{N} = \frac{1}{gm}\sum_j\sum_i x_{ji}.$

Note that the sample variance using all gm units can be expressed as

$$s_*^2 = \frac{1}{mg-1}\left[(m-1)\sum_{j=1}^{g} s_j^2 + \frac{m}{N^2}\sum_{j=1}^{g}(t_j - T)^2 - \frac{mg}{N^2}(t_{**} - T)^2\right].$$

Hence

$$E(s_*^2) = \frac{1}{mg-1}\left[g(m-1)S^2 + \frac{m}{N^2}\sum_{j=1}^{g}\mathrm{Var}(t_j) - \frac{mg}{N^2}\mathrm{Var}(t_{**})\right].$$

Rewriting this gives

$$\mathrm{Var}(t_{**}) = N^2\left(\frac{m-1}{m}\right)S^2 + \left(\frac{1}{g}\right)\sum_{j=1}^{g}\mathrm{Var}(t_j) - N^2\left(\frac{mg-1}{mg}\right)E(s_*^2).$$

Therefore, by replacing S^2 and $\mathrm{Var}(t_i)$ with unbiased estimates and replacing $E(s_*^2)$ with s_*^2, we can generate approximately unbiased estimates of $\mathrm{Var}(t_{**})$.

It may be worth emphasizing that this result does not require the user to know anything about the original sample design. If users are given a way to invert the original design, then they can, by repeated subsampling, achieve nearly the efficiency of the original design and readily estimate the appropriate sampling errors. There is one condition on this result, namely that the subsample size be such that $m \geq 2$. Incidentally, for $m = 2$, the variance expression becomes

$$\mathrm{Var}(t_{**}) = \frac{N^2}{2}S^2 + \left(\frac{1}{g}\right)\sum_{j=1}^{g}\mathrm{Var}(t_j) - N^2\left(\frac{2g-1}{2g}\right)E(s_*^2).$$

Based on this, as above, a variance estimator could be built for two PSU per stratum designs.

3.3 An SOI Illustration

In this subsection we consider an example of an inverse algorithm and how well it works. The Statistics of Income (SOI) corporate sample will be our starting point. Now, as noted earlier, the SOI sample has essentially a stratified SRS design and so can be inverted (subsection 2.2).

It is our belief that many SOI users might find a full SRS inverse sample more valuable and easier to employ than the complete, stratified sample data base. An interim goal could be to provide them with a set of simple random samples. A more flexible system would be to provide the interactive software to allow the user to designate the simple random samples of interest, to be selected from the complete data base.

In our simulations we used four of the strata in the SOI sample of corporate returns, namely the strata representing the smallest regular corporations (Hughes *et al.* 1994). As can be seen from table 1, the stratified sample (of four strata) consisted of 15,618 units, and the largest SRS that can be selected is $m = 2{,}224$. The table also shows the population sizes and the estimated variance of the variable Total Assets, within each stratum.

Table 1
Corporate Population and Sample Size, plus Estimated Stratum Variances, For Four SOI Stratum

Strata (h)	N_h	n_h	S_h^2 (in 1000's)
1	1,376,801	3,889	222,808
2	552,909	2,224	670,162
3	678,371	4,005	12,796,578
4	436,023	5,500	14,984,753

The variable total assets was used because it is the primary stratifying variable; and, therefore, the loss in precision due to removing the stratification should be relatively large. Indeed, this proved to be the case.

Shown below is the ratio of the variance of the estimated total using *g* simple random samples, of 2,224 each, divided by the variance of the total based on the stratified sample. The table displays values of *g* from 1 to 1,000. For example, if only one SRS is selected the variance of the estimated total is 29 times larger than the variance of the stratified total.

g	Relative Variance Increase
1	29.31
2	15.16
10	3.83
100	1.28
500	1.06
1000	1.03

By resampling 500 to 1,000 times, the variance has been reduced to the same order of magnitude as the stratified sample. Even at 100 subsamples good results exist here, suggesting that the use of an inverse algorithm could work well for strata such as these. This is not to recommend that an inverse algorithm be employed in general with so few resamples. Doubtless, in highly skewed populations a much larger number would be required.

4. POTENTIAL APPLICATIONS AND NEXT STEPS

In this paper we have shown that inverse sample design algorithms exist in a few special cases. We do not, as yet, have a general result – if, indeed, there is one. This is clearly a part of the problem that needs more work. Like most tools, an inverse sampling algorithm may not be the best choice in certain cases; it may not be even a reasonable alternative in some circumstances. But there are applications where it appears to have advantages and so should be considered. In this section we both briefly suggest areas where this methodology may be useful and also mention some of the limitations and problems that remain.

Customer-Driven Perspective – It is worth emphasizing the customer-driven nature of our approach. Even if it could not be justified on other grounds, inverse algorithms might be advocated as a part of "reinvention" (*e.g.*, Osborne and Gaebler 1992). Right now many large complex surveys may not be sufficiently benefiting society, because they are so badly under-analyzed or even misanalyzed:

- For the long run, we must work towards increasing the survey and general quantitative literacy of existing and potential customers – *e.g.*, as with the new series *What Is a Survey?* (Scheuren (ed.) 1995).
- In the short run, we need to start where our customers are – giving due respect to the often small part that survey data may add to their decision making. Certainly it is worth thinking about ways to lower the cognitive costs customers bear when using our complex survey "products."

A "Sample" of Possible Opportunities – There is an increasing awareness of the weaknesses within the traditional randomization paradigm (*e.g.*, Särndal and Swensson 1993). Of particular concern here is all the fiddling we have to do when trying to correct for nonsampling errors. Some of this flavour is evident in Rao and Shao (1993). By putting the possible adjustments for these nonsampling errors back into a simple random sampling framework, we may, indeed, be able to make more progress.

For decades, survey practitioners have elaborated exceedingly complex sample designs; and, then, made efficient point and confidence interval estimates from them. On the other hand, how much do we really understand about the distributions that our sample estimators generate when effective sample sizes are small to moderate? Will we be able to fully capitalize on the "visualization revolution" now occurring (*e.g.*, Cleveland 1993)? Particularly in the presence of nonsampling error? Maybe we should be building in a way to always look at distributions. The use of an inverse sampling algorithm might be one possibility (See also Pfeffermann and Nathan 1985). In any case, stronger visualization tools for complex surveys could help, even the very experienced among us, deepen our intuitions and connect them better to the particular population under study. Obviously, visualization efforts also pay off by lowering the price customers pay to use survey data.

An intriguing problem where the inverse sampling algorithm may have an application is the case where we have a two PSU per stratum design with L strata where L is small, say less than 30. Suppose further that for some of the variables in the survey the stratification and clustering are unimportant – *i.e.*, the design effect is $\delta = 1$, approximately. For these variables, would it not be possible for the stability of the variance estimate to be greater with the resampled method than with the Balanced Repeated Replication (BRR) approach to variance estimation that is usually employed?

Another example that we are considering is the case where the user is interested in tests of independence in 2 × 2 tables, based on stratified sample data (Hinkins, Oh and Scheuren 1995). For the chi-square test statistic we are now in the midst of comparing our results with the approach suggested by Scheuren (1972) and Fellegi (1980). So far it appears that the power of our method is comparable to these more familiar approaches (as might be expected from, say, Westfall and Young (1993)). This may be an instance where the extra work involved in the inverse sampling algorithm may have real benefits – beyond just making it easier for users to employ familiar tools – by allowing the user to look at the distribution rather than just one p-value.

A "Sample" of Problems Remaining – A "sample" of the problems that remain with our inverse algorithm might be given here. For example, what happens when we do not know what the population size is? What happens when the population has more than one elementary unit – persons, say, for one analysis; households for another; neighbourhoods for still a third? Answers exist for these difficulties but they have an *ad hoc* flavour to us. In many surveys, for instance, we guess about N and use that guess in poststratification. That degree of approximation for an inverse might be acceptable. For the problem of multiple analysis units, we could do several inverses. While potentially workable, this seems exceedingly awkward.

We have indicated that in some cases it may not be too difficult to resample multiple times using the inverse algorithm in order to reproduce reasonable efficiency. But what about the case where the user of a stratified sample is interested in subpopulations. If the domains of interest are in fact the strata, then the user does not gain any benefits by using the SRS's produced using the inverse algorithm. If the domains of interest cut across the strata and they are small, then the number of samples required using the inverse algorithm may be very large in order to maintain reasonable estimation for the domains.

Finally, we briefly mention one more problem that we have thought about. Many multistage designs actually select only one PSU per stratum. The strata are then paired for variance estimation purposes. We have already noted that an inverse to this approximation is available which can be made about as good as that approximation is to begin with. Is there a way to get a better approximation using the inverse approach directly?

Last Words – Many things are changing in our profession. The worldwide quality revolution certainly has had an impact (Mulrow and Scheuren 1996). We are remaking the way surveys are done – from design, to data capture, to the way customers use them. This paper may be a small contribution to that process.

ACKNOWLEDGEMENTS

We wish to express our particular appreciation to the referees and associate editor for their insightful prodding and scholarship. The original submission we sent in was only a sketch of what is now included. We also owe a debt of gratitude to Phil Kott, who has been discussing our ongoing work at various Washington Statistical Society meetings.

APPENDIX

Suppose one has a cluster sample of k clusters from a population of N clusters, where each cluster has the same number of units, M. In the inverse sampling algorithm, the first step is to choose the vector $(m_1, m_2, \dots, m_k)$ containing the number of units to be chosen from each cluster. Let q indicate the number of nonzero values of m_i. The probability of selecting the one pattern with $q = k$, that is the pattern with $m_i = 1$, for all $i = 1, 2, \dots, k$, is

$$\Pr(q = k) = M^{k-1}\frac{(N-1)(N-2)\dots(N-k+1)}{(NM-1)(NM-2)\dots(NM-k+1)}.$$

Call this probability P_1. If $NM >>> k$ then P_1 can be approximated by

$$\prod_{i=1}^{k-1}\frac{(N-i)}{N} = \frac{(N-1)(N-2)\dots(N-k+1)}{N^{k-1}}.$$

Consider next the partition of k corresponding to $q = k - 1$; this corresponds to exactly one partition of k, namely $\{1, 1, \dots, 1, 2\}$. There are $k(k-1)$ equally likely possible patterns of $(m_1, \dots, m_k)$ with $q = k - 1$. The probability of selecting a vector $\boldsymbol{m}$ with $q = k - 1$, is

$$\Pr(q = k - 1) = \frac{k(k-1)(M-1)}{2M(N-k+1)}P_1.$$

Therefore it is not difficult to calculate the probability that the selected $\boldsymbol{m}$ has either $q = k$ or $q = k - 1$. The following table shows some examples for two values of M.

Table A
$\Pr(q = k - 1 \text{ or } q = k)$

k	N	$M = 10$	$M = 100$
4	8	.92	.90
4	20	.99	.98
10	20	.38	.34
10	30	.63	.59
10	50	.83	.80
10	200	.99	.98
50	500	.35	.30
50	1000	.70	.66
50	5000	.98	.98

For small k, it is not difficult to calculate the entire probability distribution needed to generate $\boldsymbol{m}$. But as k increases, the number of partitions increases, and this calculation becomes difficult or at least tedious. For $k = 4$, there are only 4 partitions; for $k = 10$ there are 39 possible partitions. One can see from Table A, that as the cluster sample becomes "larger," if the sampling rate is small enough, *i.e.*, if $k << N$, then one might only need to calculate the probabilities for these two partitions in order to approximately invert the cluster sample. For $k = 10$ and $N = 200$, these two partitions essentially account for all of the probability distribution.

The probability of selecting just one unit per cluster ($q = k$) is smaller than the values in Table A; so, in order to use a systematic inverse, we would want $k <<< N$. This can be obtained in some settings when the number of clusters is large and we are willing to take k very small, relying on repeatedly resampling the original survey, as described in Section 3.

To illustrate, assume a sample of size k_0 where, of course, $k_0 < k$, so that an inverse is possible; Further, to see if a systematic inverse would work, let $k_0 <<< N$. This is the case we illustrate in table B. In table B, we have confined attention to just one value of N, N = 5000 clusters, although the results could be extended readily.

Table B

Pr{inverse sample picks the pattern (1,1, ..., 1)}

k_0	k_0/N	$M = 10$	$M = 100$
2	.0004	.9998	.9998
5	.001	.9982	.9980
10	.002	.9919	.9911
20	.004	.9663	.9627
30	.006	.9245	.9166
40	.008	.8687	.8553
50	.01	.8015	.7821

Clearly, as k/N gets small, a systematic sample becomes a better and better approximate inverse. Only experience would confirm if the approximation at $k_0 = 20$ and $k_0/N = .004$, say, is adequate. We think it might be, especially since the effect of using a systematic inverse usually is to make the variance calculations more conservative (since typically the intracluster correlation $\rho > 0$).

REFERENCES

BELLHOUSE, D. (1988). A brief history of random sampling methods. *Handbook of Statistics,* 6, 1-14.

CLEVELAND, W. (1993). *Visualizing Data.* Summit, NJ: Hobart Press.

COCHRAN, W. (1977). *Sampling Techniques.* New York: Wiley.

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

FELLEGI, I. (1980). Approximate tests of independence and goodness of fit based on multistage samples. *Journal of the American Statistical Association,* 75, 261-268.

HANSEN, M. (1987). Some history and reminiscences on survey sampling. *Statistical Science,* 2, 162-179.

HINKINS, S., OH, H.L., and SCHEUREN, F. (1995). Using an Inverse Algorithm for Testing of Independence Based on Stratified Samples. George Washington University Technical Report.

HUGHES, S., MULROW, J., HINKINS, S., COLLINS, R., and UBERALL, B. (1994). Section 3, *Statistics of Income – 1991, Corporation Income Tax Returns,* 9-17. Washington, DC: Internal Revenue Service.

KISH, L. (1995). The Hundred Years Wars of Survey Sampling. Centennial representative Sampling Conference, Rome, May 31, 1995.

LAHIRI, D. (1951). A method for sample selection providing unbiased ratio estimates, *Bulletin of the International Statistical. Institute,* 34, 72-86.

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. Washington, DC: Public Health Service.

MULROW, J., and SCHEUREN, F. (1996). Measuring to improve quality and productivity in a processing environment. *Data Quality,* 2, 11-20.

OSBORNE, D., and GAEBLER, T. (1992). *Reinventing Government.* New York: Plume.

PFEFFERMANN, D., and NATHAN, G. (1985). Problems in model identification based on data from complex samples. *Bulletin of the International Statistical Institute,* 68.

RAO, J.N.K., and SHAO, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika,* 79, 811-822.

RAO, J.N.K., and WU, C.F.J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association,* 83, 231-241.

SÄRNDAL, C.-E., and SWENSSON, B. (1993). Washington Statistical Society talk on the shifting nature of the survey sampling paradigm.

SCHEUREN, F. (1972). Topics in Multivariate Finite Population Sampling and Data Analysis. George Washington University Doctoral Dissertation.

SCHEUREN, F. (Ed.) (1995). *What is a Survey?* One of a series of pamphlets published by the American Statistical Association to increase survey literacy.

SKINNER, C., HOLT, D., and SMITH, T., (Eds.) (1989). *Analysis of Complex Surveys.* New York: Wiley.

WESTFALL, P., and YOUNG, S. (1993). *Resampling-Based Multiple Testing.* New York: Wiley.

WOLTER, K. (1985). *Introduction to Variance Estimation.* New York: Springer-Verlag.