INVERSE SAMPLING DESIGN ALGORITHMS

Susan Hinkins, H.Lock Oh and Fritz Scheuren
Internal Revenue Service

Fritz Scheuren, 1402 Ruffner Rd., Alex.,Va 22302 U.S.A. or Email scheuren@aol.com

KEY WORDS: Inference in Complex Surveys, IID Sampling, Resampling

1. Introduction and Background

The development of modern survey sampling is an extraordinary achievement (Bellhouse, 1988). Many of us in the U.S. federal government know or knew the early pioneers personally: Morris Hansen, Bill Hurwitz, and so many others(Hansen, 1987, Bailar, 1989). After all, much of the work was done in Washington -- starting with the lectures that Neyman gave in the 1930's when he was invited to the U.S. by W. Edwards Deming (Duncan and Shelton, 1978).

Since those early beginnings, Neyman's insights about randomization-based inference (Neyman, 1934) have, of course, been expanded and elaborated. Subtle tools now exist for a range of practical settings. It must be added, too, that concerns about the limits of the randomization paradigm have also grown; this has been so, especially in recent years, with the rise in the "respectability" of model-assisted and even full model-based inferences from surveys (e.g., Särndal, Swensson, and Wretman, 1992).

The very richness in the development of randomization-based designs 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 that the random variables being observed are obtained from a sampling process that makes them 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, et al., 1989); and much time and effort have been devoted to it in software written for surveys (e.g., Wolter, 1985).

With all that has been done already, can something more of value be added? We think we may have a small contribution to offer on how to deal better with the "seam" which currently exists between IID and survey statistics. We do not (yet) address model-based inference issues; but conjecture, nonetheless, that our approach might provide yet another viewpoint that could increase understanding of the various perspectives.

Organizationally, the paper is divided into four sections. This introduction is Section 1. In Section 2 a general problem statement is provided and a proposed "resolution" offered. A concrete illustration of our ideas is 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 and Mulrow et al, 1994). Because of space limitations, the simulations done are only covered briefly in a concluding section (Section 4). It is there, also, that we discuss a few of the many next steps needed for our still embryonic ideas to grow useful.

2. Problem Statement and Proposed "Resolution"

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?

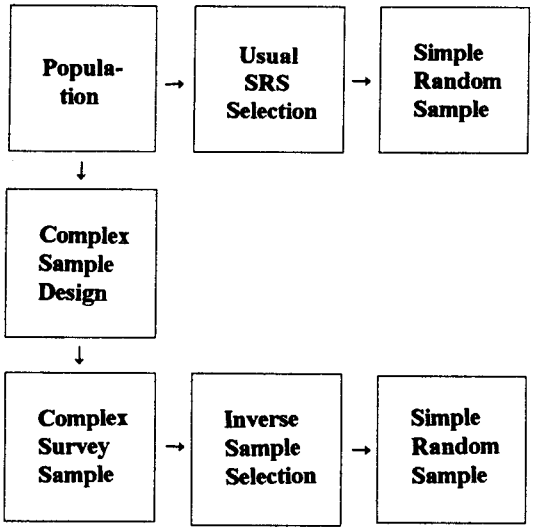
It is our contention that some of the time the answer to this question is "Yes." We call this second sample design an "Inverse Sampling Design Algorithm" -- hence, the name of this paper.

A schematic might help visualize the algorithm (see next page). 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., Cochran, 1977), such that all possible 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 IID 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.



The nature of the algorithms we are talking about should, by this point, be obvious. They 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) Apply your conventional statistical package (or perhaps model-based estimator?) 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 flavor 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. (For many of us in government this may not be true yet; but it is coming!) 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; 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.

3. An Example: A Stratified Sample

Suppose that we wish to draw a simple random sample, without replacement, from a finite population of size N. However, the population is no longer available for sampling, but we have a stratified sample, with say four strata, taken from this population. The stratified sample was taken with fixed sample sizes n_h from each stratum h, and known stratum population sizes, $N_1 + N_2 + N_3 + N_4 = N$. We need to select our simple random sample (without replacement) by resampling from this stratified sample. The largest simple random sample (SRS) that can be selected in this manner is of size $m = \min\{n_h\}$.

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, so that:

$$\Pr(m_1=i_1, m_2=i_2, m_3=i_3, m_4=i_4) = \frac{\binom{N_1}{i_1}\binom{N_2}{i_2}\binom{N_3}{i_3}\binom{N_4}{i_4}}{\binom{N}{m}}$$

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

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.

This procedure will reproduce a simple random sampling mechanism unconditionally, i.e., when taken over all possible stratified samples. That is, for any given simple random sample of size m from the original population, the probability of selecting this sample will be

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

The probability of selecting a particular SRS is equal to the probability of selecting that SRS from the stratified sample, given that the SRS is contained in the stratified sample, multiplied by the probability that this SRS is contained in the stratified sample. For this given simple random sample, 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 this SRS 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

$$\frac{\binom{N_1-x_1}{n_1-x_1}\binom{N_2-x_2}{n_2-x_2}\binom{N_3-x_3}{n_3-x_3}\binom{N_4-x_4}{n_4-x_4}}{\binom{N_1}{n_1}\binom{N_2}{n_2}\binom{N_3}{n_3}\binom{N_4}{n_4}}$$

Given that this SRS is contained in the stratified sample, the probability of selecting the SRS, using the method described, is equal to the probability of selecting the pattern (x_1,x_2,x_3,x_4), times the probability of selecting this SRS given this pattern, or

$$\frac{\binom{N_1}{x_1}\binom{N_2}{x_2}\binom{N_3}{x_3}\binom{N_4}{x_4}}{\binom{N}{m}} \times \frac{1}{\binom{n_1}{x_1}\binom{n_2}{x_2}\binom{n_3}{x_3}\binom{n_4}{x_4}}$$

Multiplying this equation times the previous one shows that the probability of selection for this simple random sample, using the proposed method of subsampling from the stratified sample, is

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

For many applications, a simple random sample is much easier to use **correctly** than a complicated stratified sample data base. However, by subsampling from the stratified sample, we lose power both by decreasing the sample size, from n to m, and by losing whatever increase in precision was due to stratification. The following discussion gives an example of the trade-offs involved; also how this loss in power might be addressed.

In the SOI environment at the Internal Revenue Service, our primary microdata users are very familiar with our sample designs and quite knowledgeable about how to use the stratified sample of corporations being drawn . But we have other users who do not use our data regularly enough to be familiar with all the design's intricacies. We know of situations where our sample, having as many as 53 strata, has been treated as if it were SRS. This can bias the estimates.

To illustrate our concerns, we examine below data taken from four of the SOI strata (those for the smallest regular corporations). In particular, we will look at the mean squared errors of four estimators of the population mean: (1) the stratified sample mean, sample size n; (2) the mean from a simple random sample, with the same total sample size, n; (3) the mean using the largest simple random sample that could be subsampled from the stratified sample, $m = \min\{n_h\}$; and, finally, (4) the sample mean when the stratified sample is incorrectly used -- as if it were from a simple random sample.

Since the first three are unbiased estimates, the mean square error is equal to the variance of the estimator. The last estimator is biased, and we use the bias squared for comparison; though this is only one component of the mean square error, it is the dominant component.

Three variables are considered: total assets, variable A (which is a stratifying variable); net income, variable B (which is one component of another stratifying variable); the third variable, C, total taxes after credits, is not used at all in the stratification.

In the following table, comparing the second row to the first (where the first has been normed to 1) shows the loss in power due to not using the stratification; the sample sizes are the same (n=15,618). Not surprisingly, the largest relative loss in power is for the primary stratifying variable, total assets. Comparing the third row to the second shows the further increase in variance due to the smaller sample size (m=2,224). These losses are more nearly the same for all variables. There is a very significant increase in the variance by using the smaller, subsampled SRS compared

to the original stratified sample (2x7, say, for net income -- variable B).

Row	A	B	C
1	1	1	1
2	4+	2-	2+
3	7-	7+	7-

Despite the above, an SRS subsample may be preferable to using the stratified sample incorrectly as a simple random sample (where the increase in mean square error is literally about 1,000 times greater).

Drawing a single, smaller simple random sample from our larger, more complex stratified sample might be enough for some of our users. However, for other users the loss in power shown between the original estimates based on the stratified sample and the simple random sample may not be acceptable.

As a means to increase the power of our approach, it was natural to consider resampling techniques. Take, for instance, the simplest case, where the user is interested in estimating means (or totals). By repeating the entire subsampling procedure, we can generate k simple random samples each of size m, where each SRS is selected independently from the given stratified sample. Each repetition must include both steps of the subsampling procedure, beginning with redrawing the stratum subsample sizes from the hypergeometric distribution.

Let $\bar{x}_{*1}$ denote the mean of one SRS of size m subsampled from the stratified sample. Let $\bar{x}_{**}$ denote the mean over all km units in the k simple random samples each of size m. Finally, let $\bar{x}_{st}$ denote the original stratified sample mean. Then, the increase in the variance using the km units, rather than the stratified sample, is

$$\mathrm{Var}(\bar{x}_{**}) - \mathrm{Var}(\bar{x}_{st}) = \{\, \mathrm{Var}(\bar{x}_{*1}) - \mathrm{Var}(\bar{x}_{st}) \,\} / k .$$

This follows since, conditional on the stratified sample, the expected value of x_{*1} is equal to the stratified sample mean, x_{st}. Because the k replications of the simple random sampling process are performed independently, given the stratified sample, then

$$\text{for } i \neq j,\quad E(\bar{x}_{*i}\bar{x}_{*j} \mid strat.\ sample) = E(\bar{x}_{*i} \mid strat) \times E(\bar{x}_{*j} \mid strat)$$

$$= (\bar{x}_{st})^2$$

Therefore, unconditionally, for i not equal to j,

$$Cov(\bar{x}_{*i}, \bar{x}_{*j}) = E(\bar{x}_{*i}\bar{x}_{*j}) - \bar{X}^2$$

$$= E_{strat}(\bar{x}_{st}^2) - \bar{X}^2$$

$$= Var(\bar{x}_{st})$$

And

$$Var(\bar{x}_{**}) = Var\left(\frac{1}{k} \times \sum_{i=1}^{k} \bar{x}_{*i}\right)$$

$$= \frac{1}{k^2}\left(\sum_{i=1}^{k} Var(\bar{x}_{*i}) + \sum_{i=1}^{k}\sum_{j \neq i} Cov(\bar{x}_{*i}, \bar{x}_{*j})\right)$$

$$= \frac{1}{k^2}\left(k\, Var(\bar{x}_{*1}) + k(k-1) Var(\bar{x}_{st})\right)$$

$$= Var(\bar{x}_{st}) + \frac{1}{k}\left(Var(\bar{x}_{*1}) - Var(\bar{x}_{st})\right)$$

(This result can be generalized to all linear functions; and, approximately, to nonlinear functions that can be linearized by a Taylor series.)

To give numeric content to the above, consider the variable total assets in the previous example, where the original stratified sample variance again has been normed to 1 . The following shows the normed variances of the sample means based on k simple random samples of 2,224 each, for increasing values of k:

k	Var($\bar{x}$)
1	29.31
2	15.16
10	3.83
100	1.28
500	1.06
1000	1.03

By resampling 500 to 1000 times, the variance has been reduced to the same order of magnitude as the stratified sample. Even at 100 subsamples good results exist here (an insight we employed in our simulations, as mentioned in Section 4).

Many SOI users, familiar as they are with IID statistical methods, would find an SRS more valuable and easier to employ, than our complete, stratified sample data base. An interim goal might 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.

4. Additional Considerations and Next Steps

This section concludes with several short topics. First, a little more theory is given in connection with Section 3 -- in particular, how to estimate the variance without direct knowledge of the original sampling design. Second, some of our many simulation results are covered. Finally, there are discussions of next steps, plus concluding comments -- among them an invitation for possible joint work.

A Little More Theory.--Let S^2 and X denote the population variance and population mean for the variable X. For the sample means and variances calculated from the generated simple random samples, let

$$\bar{x}_{**}=\frac{1}{k}\sum_{j=1}^{k}\bar{x}_{j*}=\left(\frac{1}{km}\right)\sum_{j=1}^{k}\sum_{i=1}^{m}x_{ji}$$

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

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

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

$$s_*^2=\left(\frac{1}{mk-1}\right)\left[(m-1)\sum_{j=1}^{k}s_j^2+m\sum_{j=1}^{k}(\bar{x}_{j*}-\bar{X})^2-mk(\bar{x}_{**}-\bar{X})^2\right]$$

Hence

$$E(s_*^2)=\left(\frac{1}{mk-1}\right)\left[(m-1)kS^2+m\sum_{j=1}^{k}Var(\bar{x}_j)-mkVar(\bar{x}_{**})\right]$$

Rewriting this gives

$$Var(\bar{x}_{**})=\left(\frac{m-1}{m}\right)S^2+\left(\frac{1}{k}\right)\sum_{j=1}^{k}Var(\bar{x}_j)-\left(\frac{mk-1}{mk}\right)E(s_*^2)$$

Therefore, by replacing S^2 and $Var(\bar{x}_j)$ with unbiased estimates and replacing $E(s_*^2)$ with s_*^2, we can generate unbiased estimates of $Var(\bar{x}_{**})$. **This result does not require the user to know anything about the original sample design.**

Some Simulation Results.-- An extensive series of simulations were conducted as part of our work on this problem. Space only permits a brief summary:

(1) Pseudo-populations Created.--A version of the estimation problem set out in Section 2 was studied for n=156 and m=22. To do this, we generated a population of 3,044 Multivariate normal observations

z' = (total assets, net income, tax after credits)

with the same means, variances, and strata definitions as in the SOI corporate population.

(2) Estimation Research.--Repeated stratified samples were selected (10 in all). Then, from each of these, 100 subsamples were drawn for study. To accompany the two-step SRS sampling, 1,000 one-step SRS samples were also drawn for comparison purposes. Quantile-quantile charts were employed in the analysis and these showed the expected agreement between the two SRS methods for the sample mean. (For this case, a direct comparison with the stratified sample is readily available, as has been seen.)

(3)Hypothesis Testing.-- To accompany the estimation simulations discussed above, 2x2 tables were constructed from the same samples, to look at the relationship between total assets and net income. Each variable was split at the median; hence, under the null hypothesis of independence, the expected cell sizes were all 5.5. Both a chi-square and a Fisher exact test were conducted. Again, the one- and two-step SRS results agreed in distribution.

(4) Initial Comments on Simulation.-- For Fisher's exact test, no readily available alternative exists in the stratified case -- so we are looking at an instance where the extra work involved in the two-step sampling may have real benefits, beyond just making it easier for users to employ familiar tools. 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). Our belief is that the power of our method will equal or exceed these more familiar approaches.

Next Steps.- - At best, in this paper we have done no more than shown that an inverse sample design algorithm exists in one limited setting -- that of stratified sampling. What about cluster sampling? multistage designs? and on and on?

It would be great to be able to say that for these other (more?) interesting surveys that we have worked out general inverses or have a way to characterize when an inverse design exists(even approximately). At this point,

though, all we have are a few hunches about how to "invert" some of the more common designs. Instead of covering these, however, it may make sense to connect up what we have been talking about with some of the other problems we have as samplers and as government statisticians.

First, 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 sample surveys may not be sufficiently benefiting society, because they are so badly underanalyzed or even misanalyzed . Of course, we must work towards increasing the survey and other quantitative literacies of existing and potential customers. Nonetheless, for the short run, we need to start where they are -- giving due respect to the small part that survey data may add to their decisionmaking. Certainly it is worth thinking about ways to lower the cognitive costs customers bear when using our "products."

Second, 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. 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 (e.g., on the current multiple imputation controversy).

Third, many in this audience have done exceedingly complex sample designs and made elaborately efficient estimates from them. On the other hand, how much do we really understand about the distributions that our sample estimators generate? Will we be able to fully capitalize on the "visualization revolution" now occurring (Cleveland, 1993)? particularly in the presence of nonsampling error? Maybe we should be building in a way to **always** look at distributions? This could help even the very experienced among us deepen our intuitions and connect them better to the particular population under study.

Concluding Comments.--Many things are changing in our profession. 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 the paradigm shifts underway. We hope so.

Obviously, we have a lot more to do to develop the ideas presented here today. Please consider joining us by looking at inverse algorithms for your own surveys and comparing the results from them with existing methods of analysis. It is likely that taking up this challenge could lead to some very tough problems; on the other hand, it could be great fun too!

References

Bailar, B.(1989), Contributions to statistical methodology from the federal government, *Sequicentennial Invited Paper Sessions, Proceedings of the American Statistical Association*, 515-519.

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

Cleveland, W.(1994), *Visualizing Data*, Summit, NJ: Hobart Press.

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

Duncan, J. and Shelton, W.(1978), *Revolution in U.S. Government Statistics, 1926-1976*, U.S. Department of Commerce, Washington.

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.

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.

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.

Neyman, J.(1934),On the two aspects of the representative method: The method of stratified sampling and the method of purposive selection, *Journal of the Royal Statistical Society B*, 97, 558-625.

Osborne, D. and Gaebler, T. (1992), *Reinventing Government*, New York: Plume.

Rao, J. and WU, J.(1988), Resampling inference from with complex survey data, *Journal of the American Statistical Association*, 83, 231-241.

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

Särndal, C.-E., Swensson, B., and Wretman, J.(1992), *Model Assisted Survey Sampling*, New York: Springer-Verlag.

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

Skinner, C., Holt, D., and Smith, T.(eds.)(1989), *Analysis of Complex Surveys*, New York, Wiley.

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