

## UNDOING COMPLEX SURVEY DATA STRUCTURES: SOME THEORY OF INVERSE SAMPLING

J.N.K. Rao<sup>1</sup> and A.J. Scott<sup>2</sup>

### ABSTRACT

Application of classical statistical methods to data from complex sample surveys without making allowance for the survey design features can lead to erroneous inferences. Methods have been developed that account for the survey design, but these methods require additional information such as survey weights, design effects or cluster identification for micro data. Inverse sampling (Hinkins et al., 1997) provides an alternative approach by undoing the complex survey data structures so that standard methods can be applied. Repeated subsamples with simple random sampling structure are drawn and each subsample analysed by standard methods and combined to increase the efficiency. This method has the potential to preserve confidentiality of micro data, although computer-intensive. We present some theory of inverse sampling and explore its limitations.

Key Words: Confidentiality; Repeated Subsampling; Survey Design.

### RÉSUMÉ

L'application de méthodes statistiques à des données complexes sans tenir compte des caractéristiques de design du sondage peut mener à des inférences erronées. Certaines méthodes pour tenir compte du design ont été proposées, mais ces méthodes requièrent des informations additionnelles telles que les pondérations du sondage, les effets du design, ou l'identification des grappes. L'échantillonnage inverse (Hinkins et al., 1997) est une approche alternative dans laquelle les structures de données sont démontées afin de permettre l'application de méthodes statistiques standards. Plusieurs sous-ensembles comportant une structure d'échantillonnage simple sont d'abord tirés. Puis, chaque sous-ensemble est analysé par des méthodes standards, et ensuite combiné pour augmenter l'efficacité. Cette méthode a le potentiel de préserver la confidentialité des données, mais est aussi très intensive. Lors de cet exposé, nous présenterons des extraits de la théorie sur l'échantillonnage inverse et discuterons des limitations et des applications de cette méthode.

Mots Clé: Confidentialité; Plusieurs sous-ensembles; Design du sondage.

### 1. INTRODUCTION

There is a fairly clear distinction between the focus of traditional sample survey methodology and that of the rest of applied statistics. Survey samplers have concentrated on developing efficient (but complicated) ways of drawing samples to estimate rather simple quantities (population means, proportions, totals, etc.). Most other applied statisticians, by contrast, have concentrated on developing sophisticated methods for fitting very complicated models, but assuming a rather simple sampling structure (often that the observations are independent).

In reality, data from complicated surveys are often used to fit complicated models. For example, people may want to use data from a Labour Force Survey to find out what effect education has on unemployment levels. They might want to use data from health surveys to find out what effect housing conditions or poverty has on morbidity, and so on. Extending the range of application of standard methods so that they can be applied to data from complicated sample surveys, involving multi-stage sampling and variable selection probabilities, is notoriously difficult and cumbersome; see e.g., Skinner, Holt and Smith (1989).

How do practitioners deal with the complexity of survey data structures? Adapting a quote from

---

<sup>1</sup> J.N.K. Rao, School of Mathematics and Statistics, Carleton University, Ottawa (Ontario), Canada K1S 5B6, jrao@math-carleton.ca

<sup>2</sup> A.J. Scott, Department of Statistics University of Auckland, Auckland, New Zealand, scott@stat.auckland.ac.nz

Hinkins, Oh and Scheuren (1997) (abbreviated HOS hereafter): “If your only tool is a hammer, every problem looks like a nail!”; the hammer available to most people is one of the big statistical packages (SAS, Splus, SPSS, etc). Most people still just push their data through a standard program and ignore the survey design features. This is in spite of the fact that a great deal of effort over the last two decades has been spent on developing methods to analyze survey data that take account of design features, and specialized programs such as SUDAAN or WesVar are now available to implement some of these methods.

An alternative to developing complex new tools (which may rarely be used in practice anyway!) is to work backwards: instead of tailoring the methods to fit the data, tailor the data to fit the methods. One approach along these lines was developed in Rao & Scott (1992; 1999). Another approach has been suggested in HOS. Their basic idea is to avoid the pain caused by a complicated sample by choosing a subsample that has a simple random sample structure (or at least has a structure that is considerably simpler to handle than the original sample). Obviously this involves some loss in efficiency, especially if the subsample is very much smaller than the original sample, as often turns out to be necessary. However, we can increase the power by repeating the process independently many times and averaging the results.

Is it possible to produce subsamples with the desired properties? The answer is often “yes”, although the resulting subsample size,  $m$ , might have to be small (in fact, no more than  $m=2$  for some standard designs). HOS give algorithms for producing simple random samples for a number of standard designs. In this paper we look at some of the properties of the repeated subsampling procedure. We develop some basic theory in Section 2, and illustrate some of the strengths and weaknesses of the procedure in Sections 3 and 4.

We note that biostatisticians trying to analyze clustered biological data face some of the same problems as survey samplers. As has often happened before, similar ideas have been suggested independently in the sample survey and biostatistics literatures. In particular, Hoffman and Weinberg (1998) suggested simplifying clustered data by choosing one observation at random from each cluster, thus producing a set of independent observations that can be handled by standard programs, and then repeating the process many times.

## 2. BASIC PROPERTIES

The results in this section are quite general and apply equally to sample surveys and the type of clustered situation considered by Hoffman and Weinberg (1998). Suppose that we are interested in estimating some population parameter,  $\theta$  say, and we have a sample,  $s_0$ , of observations drawn from the population according to some complex design. We assume that we have a subsampling algorithm that can produce samples from some simpler design. This design will often be simple random sampling, but we can extend the range of applications considerably by allowing for the possibility of more general (sub-)designs. Our only requirement for the simpler design is that we can produce an estimate,  $\hat{\theta}$ , of the quantity of interest, together with an estimate,  $\hat{V}(\hat{\theta})$ , of its variance. Let  $\hat{\theta}_j^*$  and  $\hat{V}_j^*$  denote the estimates produced from the  $j$ th subsample when we generate a sequence of  $g$  independent subsamples  $s_j^*$ . (Note that the  $\hat{\theta}_j^*$ s are not unconditionally independent when averaged over the distribution of the initial sample,  $s_0$ .) Our estimate is

$$\hat{\theta}_g = \frac{1}{g} \sum_{j=1}^g \hat{\theta}_j^*. \quad (1)$$

### Theorem 1

1. Conditional on the original sample,  $s_0$ ,  $\hat{\theta}_g$  converges almost surely to  $E\{\hat{\theta}_1^* | s_0\} = \hat{\theta}_\infty$ , say, as  $g \rightarrow \infty$ .
2.  $E\{\hat{\theta}_g\} = E\{\hat{\theta}_1^*\}$ .
3.  $Var\{\hat{\theta}_g\} = Var\{\hat{\theta}_\infty\} + \frac{1}{g} E\{Var\{\hat{\theta}_1^* | s_0\}\}$ .
4. If  $r_g = \frac{Var\{\hat{\theta}_g\}}{Var\{\hat{\theta}_\infty\}}$ , then  $r_g = 1 + \frac{r_1 - 1}{g}$ .

### Proof

1: Follows directly from (1) on noting that, conditional on  $s_0$ ,  $\hat{\theta}_1^*$ ,  $\hat{\theta}_2^*$ , . . . ,  $\hat{\theta}_g^*$  are i.i.d. (bounded) random variables.

2: Follows from the standard relationship between conditional and unconditional expectations:

$$E\{\hat{\theta}_g\} = E\{E\{\hat{\theta}_g | s_0\}\} = E\{E\{\frac{1}{g} \sum_{j=1}^g \hat{\theta}_j^* | s_0\}\} = E\{\hat{\theta}_1^*\}.$$

3: Follows from the corresponding result for variances, and the conditional independence of the  $\hat{\theta}_j^*$ s given  $s_0$ :

$$\begin{aligned} \text{Var}(\hat{\theta}_g) &= \text{Var}\{E(\hat{\theta}_g | s_0)\} + E\{\text{Var}(\hat{\theta}_g | s_0)\} \\ &= \text{Var}(\hat{\theta}_\infty) + (1/g)E\{\text{Var}(\hat{\theta}_1^* | s_0)\} \end{aligned}$$

4: Follows directly from 3.

Result 4 of Theorem 1 demonstrates that increasing the number of subsamples does indeed increase efficiency. More precisely, the bias stays constant while the variance has the form  $a + b/g$ . It also demonstrates a severe limitation of the subsampling method. If the subsample estimator,  $\hat{\theta}_1^*$ , is unbiased, then so is the resampling estimator,  $\hat{\theta}_g$ . However, if  $\hat{\theta}_1^*$  has bias of order  $\frac{1}{m}$ , where  $m$  denotes the subsample size, then  $\hat{\theta}_g$  has exactly the same bias. Since  $m$  will usually be very much smaller than the original sample size, this bias can be appreciable.

Result 4 of Theorem 1 can be used to decide how many samples are needed to obtain reasonable efficiency. For example, HOS give an example in which  $r_1 = 29.3$ . The original sample was a very efficient stratified random sample with  $n = 15,618$  observations taken from the Statistics of Income corporate survey, while the subsample was a simple random sample of  $m = 2,224$  observations, so a single subsample is relatively very inefficient. However, in this case, repeated resampling recovers all the information in the original sample in the limit. Applying Result 4 of Theorem 1 leads immediately to the following table:

$g$	1	10	100	1000
$r_g$	29.3	3.83	1.28	1.03

(HOS produced these same results by simulation but this is unnecessary in view of Result 4 of Theorem 1.) We see that  $g = 100$  subsamples would be adequate for many purposes and that we obtain almost full efficiency with  $g=1000$ .

The fact that  $\hat{\theta}_1^*, \hat{\theta}_2^*, \dots, \hat{\theta}_g^*$  are not unconditionally independent means that estimating  $\text{Var}\{\hat{\theta}_g\}$  is not completely straightforward. However, Theorem 2 leads to a relatively simple variance estimator.

## Theorem 2

$$\text{Var}\{\hat{\theta}_g^*\} = \text{Var}\{\hat{\theta}_1^*\} - \frac{g-1}{g} E\{\text{Var}\{\hat{\theta}_1^* | s_0\}\} \quad (2)$$

### Proof

Theorem 2 follows from applying Result 3 of Theorem 1 with  $g = 1$  to obtain

$$\text{Var}\{\hat{\theta}_\infty\} = \text{Var}\{\hat{\theta}_1^*\} - E\{\text{Var}\{\hat{\theta}_1^* | s_0\}\}$$

and then substituting this expression for  $\text{Var}\{\hat{\theta}_\infty\}$  in Result 3 of Theorem 1 for general  $g$ .

We can estimate the first term of (2) by  $\hat{V}_j^*$  for  $j=1, 2, \dots, g$ , and hence by their average  $(1/g)\sum_1^g \hat{V}_j^*$ . In addition, the quantity  $\sum_1^g (\hat{\theta}_j^* - \hat{\theta}_g)^2 / (g-1)$  gives an unbiased estimator of  $E\{\text{Var}\{\hat{\theta}_1^* | s_0\}\}$  since  $\hat{\theta}_1^*, \hat{\theta}_2^*, \dots, \hat{\theta}_g^*$  are conditionally independent given the initial sample,  $s_0$ . This leads to an estimator of the form

$$\hat{V}_g = \frac{1}{g} \sum_1^g \hat{V}_j^* - \frac{\sum_1^g (\hat{\theta}_j^* - \hat{\theta}_g)^2}{g} \quad (3)$$

for  $\text{Var}\{\hat{\theta}_g\}$ . The properties of the variance estimator  $\hat{V}_g$  depend on the properties of the subsample estimator  $\hat{V}_j^*$ . For example, if  $\hat{V}_j^*$  is unbiased, then  $\hat{V}_g$  is also unbiased. It may be noted that  $\hat{V}_g$  can take negative values, especially for small  $g$ .

Some obvious questions remain to be answered. In particular, it is natural to ask whether or not repeated sampling recaptures all the information in the original sample. In the stratified example given by HOS, for example, the resampling estimator (1) is fully efficient in the limit. However, it is clear from Result 2 of Theorem 1 that this is not true in general, since estimators of complex quantities like regression coefficients will usually be biased in small samples and we have seen that resampling does nothing to reduce the bias. Even when we start with an unbiased estimator, there is no guarantee that we converge to the original full sample estimator as  $g \rightarrow \infty$ .

## 3. ESTIMATION OF A TOTAL

As remarked in Section 2, the fact that resampling increases efficiency does not necessarily mean that we converge to the original full sample estimator as  $g \rightarrow \infty$ , even when we start with an unbiased

estimator for the subsample. Suppose that we start with the Horvitz-Thompson (H-T) estimator of a total for the induced (subsample) design. Theorem 3 establishes conditions under which the resampling estimator converges to the H-T estimator for the full design.

### Theorem 3

Let  $\tilde{\pi}_i(s_0)$  denote the conditional probability that the  $i$ th unit is selected in the subsample for a given initial sample,  $s_0$ . Suppose that  $\hat{\theta}_j^* = \hat{Y}_j^*$  is the H-T estimator of a total  $\theta = Y$  for the  $j$ th subsample. Then the limiting resampling estimator,  $\hat{\theta}_\infty^* = \hat{Y}_\infty^*$ , will be the H-T estimator for the original design if and only if the conditional inclusion probabilities  $\tilde{\pi}_i(s_0)$  are constant for all  $s_0$  containing the  $i$ th unit, i.e.,  $\tilde{\pi}_i(s_0) = \tilde{\pi}_i$  for all  $s_0 \supset i$ .

### Proof

We have  $\hat{Y}_j^* = \sum_{s_j^*} \frac{y_i}{\pi_i^*} = \sum_{s_0} \frac{y_i I_{ij}^*(s_0)}{\pi_i^*}$ , where  $I_{ij}^*(s_0)$  takes the value 1 if the  $i$ th unit is included in the  $j$ th subsample  $s_j^*$  and 0 otherwise, and  $\pi_i^*$  is the corresponding (unconditional) selection probability. Thus

$$\hat{Y}_\infty^* = E\{\hat{Y}_1^* | s_0\} = \sum_{s_0} \frac{y_i \tilde{\pi}_i(s_0)}{\pi_i^*}.$$

This is equal to  $\hat{Y} = \sum_{s_0} \frac{y_i}{\pi_i}$ , the H-T estimator for the

original design, if and only if  $\tilde{\pi}_i(s_0) = \frac{\pi_i^*}{\pi_i} = \tilde{\pi}_i$ , say,

where  $\pi_i$  is the probability that unit  $i$  is in the original sample.

The condition  $\tilde{\pi}_i(s_0) = \tilde{\pi}_i$  is a fairly natural one for most sampling designs for which the H-T estimator is used. It is satisfied in all but one of the algorithms suggested in HOS, for example. Apart from the single exception, their subsamples are all simple random samples of fixed size. As a result the H-T estimator for the subsample is simply the sample mean, which is the natural estimator. The exception is cluster sampling with clusters of varying sizes. In this case, it does not seem possible to obtain a simple random sample of fixed size by subsampling. Instead, we first force all clusters to have the same size by adding an appropriate number of pseudo-units to bring them up to the size of the largest cluster. Then we take one unit

at random from each sampled cluster, and discard any pseudo-units to obtain the final subsample. It is easily seen that the conditions of Theorem 3 are not met here since the conditional probability of selecting the  $i$ th unit is equal to  $\frac{1}{M(s_0)}$ , where  $M(s_0)$  is the size of the largest sampled cluster, and this clearly depends on the initial sample,  $s_0$ . In this case, however, we would not normally use the H-T estimator for the full design, preferring the ratio-to-size estimator instead. Unfortunately, there does not seem to be any simple way to obtain the ratio-to-size estimator directly by subsampling.

We might also wonder how the variance estimator  $\hat{V}_g$  compares to the variance estimator we would normally use for the original full sample. The following result gives a partial answer to this question.

### Theorem 4

If  $\hat{V}_j^*$  is the Horvitz-Thompson (H-T) variance estimator of  $\hat{Y}_j^*$  for the  $j$ -th subsample, then conditional on  $s_0$ ,  $\hat{V}_g$  converges to the Horvitz-Thompson (H-T) variance estimator of  $\hat{Y}$  for the original design, as  $g \rightarrow \infty$ , if the conditional joint inclusion probabilities are constant for all  $s_0$  containing a given pair  $(i, l)$  of units, i.e.  $\tilde{\pi}_{il}(s_0) = \tilde{\pi}_{il}$  for all  $s_0 \supseteq \{i, l\}$ .

### Proof.

The H-T variance estimator of  $\hat{Y}_j^*$  is given by

$$\hat{V}_j^* = \sum_{i, l \in s_j^*} \sum \frac{(\pi_{il}^* - \pi_i^* \pi_l^*)}{\pi_i^* \pi_l^* \pi_{il}^*} y_i y_l$$

where  $\pi_{il}^*$  is the unconditional probability that units  $i$  and  $l$  are both in the sample ( $\pi_{il}^* = \pi_i^* \pi_l^*$ ); see Cochran (1977, p. 261). Similarly, the H-T variance estimator of the full sample estimator  $\hat{Y}$  is

$$\hat{V} = \sum_{i, l \in s_0} \sum \frac{(\pi_{il} - \pi_i \pi_l)}{\pi_i \pi_l \pi_{il}} y_i y_l.$$

Also, conditional on  $s_0$  it follows from (3) that,  $\hat{V}_g$  converges almost surely to

$$\hat{V}_\infty = E(\hat{V}_1^* | s_0) - \text{Var}(\hat{Y}_1^* | s_0) \quad (4)$$

as  $g \rightarrow \infty$ . Now, noting that  $\tilde{\pi}_{il}(s_0) = \tilde{\pi}_{il} = \pi_{il}^* / \pi_{il}$ , we get

$$\begin{aligned}
E(\hat{V}_1^* | s_0) &= \sum_{i,l \in s_0} \sum \frac{(\pi_{il}^* - \pi_i^* \pi_l^*)}{\pi_i^* \pi_l^* \pi_{il}^*} \tilde{\pi}_{il} y_i y_l \\
&= \sum_{i,l \in s_0} \sum \left( \frac{\tilde{\pi}_{il}}{\pi_i^* \pi_l^*} - \frac{1}{\pi_{il}^*} \right) y_i y_l. \quad (5)
\end{aligned}$$

Further, from Cochran (1977, p. 260),

$$\begin{aligned}
\text{Var}(\hat{Y}_1^* | s_0) &= \sum_{i,l \in s_0} \sum \frac{(\tilde{\pi}_{il} - \tilde{\pi}_i \tilde{\pi}_l)}{\pi_i^* \pi_l^*} y_i y_l \\
&= \sum_{i,l \in s_0} \sum \left( \frac{\tilde{\pi}_{il}}{\pi_i^* \pi_l^*} - \frac{1}{\pi_i \pi_l} \right) y_i y_l. \quad (6)
\end{aligned}$$

It now follows from (4), (5) and (6) that  $\hat{V}_\infty = \hat{V}$ .

**Note:** With the exception of cluster sampling with unequal cluster sizes, the algorithms suggested in HOS ensure that  $\Pr(s^* | s_0)$  is the same for all  $s_0$  containing  $s^*$  which in turn imply that  $\tilde{\pi}_{il}(s_0) = \tilde{\pi}_{il}$  for all  $s_0 \supseteq \{i, l\}$ .

In Theorem 4 we considered the H-T variance estimator. But the Sen-Yates-Grundy(S-Y-G) variance estimator is often preferred over the H-T variance estimator because it is more stable and several designs for which it is always nonnegative are known, while  $\hat{V}$  frequently takes negative values (Cochran, 1977, p. 261). The S-Y-G variance estimator exists for fixed sample size designs and it is given by

$$\tilde{V} = \sum_{i,l < s_0} \sum \frac{(\pi_i \pi_l - \pi_{il})}{\pi_{il}} \left( \frac{y_i}{\pi_i} - \frac{y_l}{\pi_l} \right)^2 \quad (7)$$

for the full sample estimator  $\hat{Y}$ . Similarly, the S-Y-G variance estimator of  $\hat{Y}_j^*$  is

$$\tilde{V}_j^* = \sum_{i,l < s_0} \sum \frac{(\pi_i^* \pi_l^* - \pi_{il}^*)}{\pi_{il}^*} \left( \frac{y_i}{\pi_i^*} - \frac{y_l}{\pi_l^*} \right)^2. \quad (8)$$

Equation (4) is changed to

$$\tilde{V}_\infty = E(\tilde{V}_1^* | s_0) - \text{Var}(\hat{Y}_1^* | s_0), \quad (9)$$

where

$$\text{Var}(\hat{Y}_1^* | s_0) = \sum_{i,l < s_0} \sum (\tilde{\pi}_i \tilde{\pi}_l - \tilde{\pi}_{il}) \left( \frac{y_i}{\pi_i^*} - \frac{y_l}{\pi_l^*} \right)^2, \quad (10)$$

provided the subsample size is also fixed (Cochran, 1977, p. 260). Further,

$$E(\tilde{V}_1^* | s_0) = \sum_{i,l < s_0} \sum \frac{(\pi_i^* \pi_l^* - \pi_{il}^*)}{\pi_{il}^*} \left( \frac{y_i}{\pi_i^*} - \frac{y_l}{\pi_l^*} \right)^2. \quad (11)$$

It now follows from (9), (10) and (11) that

$$\tilde{V}_\infty = \sum_{i,l < s_0} \sum \frac{(\pi_i \pi_l - \pi_{il})}{\pi_{il}} \tilde{\pi}_i \tilde{\pi}_l \left( \frac{y_i}{\pi_i \tilde{\pi}_i} - \frac{y_l}{\pi_l \tilde{\pi}_l} \right)^2. \quad (12)$$

Comparing (7) and (12), we see that  $\tilde{V}_g = \sum \tilde{V}_j / g$  does not converge to the S-Y-G variance estimator.

If the subsample is a simple random sample unconditionally, i.e.,  $\pi_i^* = m/N$  where  $m$  is the subsample size, then  $\hat{V}_j^* = \tilde{V}_j^*$  and  $\tilde{V}_\infty = \hat{V}_\infty = \hat{V}$ , where  $\hat{V}$  is the H-T variance estimator of  $\hat{Y}$ .

Although it is reassuring to know that the resampling estimator will converge to the standard estimator for the full design under reasonable conditions, the real test comes when there is no standard estimator available for the full design. Unfortunately, even under simple random sampling, estimators of most complex parameters such as regression coefficients are biased, with a bias that is inversely proportional to the sample size. In these cases, Result 2 of Theorem 1 means that the resampling procedure can only produce estimators with bias inversely proportional to the subsample size. Since the subsample size is required to be small with most designs (no more than the number of primary sampling units in the smallest stratum in a standard stratified two-stage design, for example), this puts severe limitations on the usefulness of the method. In addition, the small subsample size often causes problems with convergence when fitting complex models involving many parameters. For example, Scott and Wild (1999) used the technique to fit logistic regression models with about ten independent variables to data from a case-control study in which controls were selected by two-stage sampling. There were approximately 300 primary sampling units, with the number of sampled individuals in a primary sampling ranging from one to six. The subsamples contained about 60 controls and the logistic regression program failed to converge in about 40% of the subsamples.

Many complex statistics can be expressed as the solution of an estimating equation. One possible way of getting round both the difficulties above is to work with the corresponding unbiased estimating equation

and to resample the equation rather than the statistic itself. This extension will be reported in a separate paper.

#### 4. CONCLUDING REMARKS

Theorems 3 and 4 showed that the resampling H-T estimator  $\hat{Y}_g$  of a total  $Y$  and its H-T variance estimator  $\hat{V}_g$  converge to the full sample H-T estimator  $\hat{Y}$  and associated H-T variance estimator  $\hat{V}$  as  $g \rightarrow \infty$ . But the properties of  $\hat{V}_g$  for finite  $g$  remains to be investigated.

#### REFERENCES

- Cochran, W. G. (1977). *Sampling Techniques*. New York: Wiley.
- Hinkins, S., Oh, H. L. and Scheuren, F. (1997). Inverse Sampling designs algorithms. *Survey Methodology*, 23, 11-21.
- Hoffman, E. and Weinberg, C. (1998). Within cluster Sampling. Paper presented at the American Statistical Association Meetings.
- Rao, J. N. K. and Scott, A. J. (1992). A simple method for the analysis of clustered binary data. *Biometrics*, 48, 577-585.
- Rao, J. N. K. and Scott, A. J. (1999). A simple method for analyzing overdispersion in clustered Poisson data. *Statistics in Medicine*, 18, 1373-1385.
- Scott, A. J. and Wild, C. J. (1999). Complex sampling and case control studies. *Bulletin of the International Statistical Institute*, 58, 327-330.
- Skinner, C. J., Holt, D. and Smith, T. (Eds.) (1989). *Analysis of Complex Surveys*. New York: Wiley.