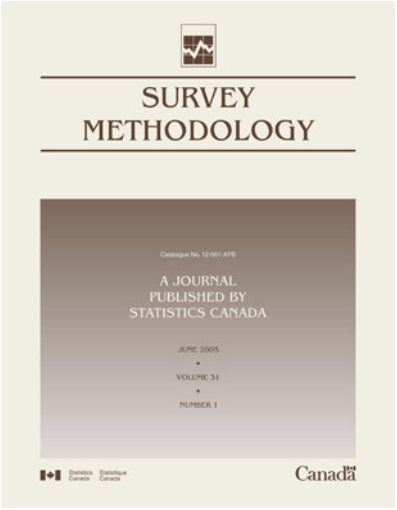
Catalogue no. 12-001-XIE

Survey Methodology

December 2003



Statistics Canada Statistique Canada

Canada

How to obtain more information

Specific inquiries about this product and related statistics or services should be directed to: Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone: 1 800 263-1136).

For information on the wide range of data available from Statistics Canada, you can contact us by calling one of our toll-free numbers. You can also contact us by e-mail or by visiting our website.

National inquiries line	1 800 263-1136
National telecommunications device for the hearing impaired	1 800 363-7629
Depository Services Program inquiries	1 800 700-1033
Fax line for Depository Services Program	1 800 889-9734
E-mail inquiries	infostats@statcan.ca
Website	www.statcan.ca

Information to access the product

This product, catalogue no. 12-001-XIE, is available for free. To obtain a single issue, visit our website at www.statcan.ca and select Our Products and Services.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner and in the official language of their choice. To this end, the Agency has developed standards of service that its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll free at 1 800 263-1136. The service standards are also published on www.statcan.ca under About Statistics Canada > Providing services to Canadians.

Statistics Canada
Business Survey Methods Division

Survey Methodology

December 2003

Published by authority of the Minister responsible for Statistics Canada

February 2006

Catalogue no. 12-001-XIE
ISSN 1492-0921

Frequency: semi-annual

Ottawa

Cette publication est disponible en français sur demande (nº 12-001-XIF au catalogue).

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued cooperation and goodwill.

Vol. 29, No. 2, pp. 177-180
Statistics Canada, Catalogue No. 12-001

Estimation of a Measure of Disclosure Risk for Survey Microdata Under Unequal Probability Sampling

C.J. Skinner and R.G. Carter [1]

Abstract

Skinner and Elliot (2002) proposed a simple measure of disclosure risk for survey microdata and showed how to estimate this measure under sampling with equal probabilities. In this paper we show how their results on point estimation and variance estimation may be extended to handle unequal probability sampling. Our approach assumes a Poisson sampling design. Comments are made about the possible impact of departures from this assumption.

Key Words: Confidentiality protection; Finite population inference; Poisson sampling; Statistical disclosure control; Uniqueness.

1. Introduction

Microdata files of survey data may be of great analytic value to researchers. When deciding whether and how to make such files available, agencies conducting surveys need to protect against risks of possible statistical disclosure (Willenborg and de Waal 2001). Skinner and Elliot (2002, abbreviated henceforth to SE) proposed a simple measure of statistical disclosure risk for survey microdata, for use as evidence to inform such decisions. They showed that this measure may be estimated simply under sampling with equal probabilities. In this paper we show how their results may be extended to handle unequal probability sampling.

The measure is introduced in section 2. Point estimation and variance estimation for the measure are considered in sections 3 and 4 respectively. See SE for the relation of this measure to the literature on statistical disclosure risk.

2. The Measure of Disclosure Risk

We consider the possible release of a microdata file consisting of a set of records for units (*e.g.*, individuals or households) in a sample *s*, selected by probability sampling from a population *U*. Each record consists of a vector of values of a specified set of variables for the given unit. Following a standard approach to disclosure risk assessment (*e.g.*, Bethlehem, Keller and Pannekoek 1990), we suppose that an intruder attempts to match the microdata records to known population units using a specified subset of variables. We assume that these "identifying variables" are categorical and that the possible combinations of their values define the categories $1, \ldots, J$ of a variable *X*. (*J* will usually be very large).

We suppose that the intruder is able to determine the value of *X* for a population unit with known identity and that the intruder 'claims' that a microdata record has been identified if and only if this value matches the value of *X* recorded in the microdata for *just one* microdata record. Assuming (a) that the population unit with known identity is randomly drawn from *U* with equal probabilities and (b) that the value of *X* for this unit is measured in the same way that *X* is measured in the microdata, the probability that the intrude's claim is correct is:

$$\theta = \Pr(\text{correct match} \mid \text{unique match}) = \sum_{j=1}^{J} I(f_j = 1) \Big/ \sum_{j=1}^{J} F_j I(f_j = 1),$$

where f_j and F_j are the frequencies of units in *s* and *U* respectively, for which $X = j$ and where $I(.)$ is the indicator function ($I(A) = 1$ if *A* is true, $I(A) = 0$ otherwise). The numerator of θ is the number of microdata records which are unique in the microdata with respect to *X* and the denominator of θ is the number of units in the population which share the same value of *X* with any of these records.

The quantity, θ, is the measure of disclosure risk considered in this paper. To protect against disclosure, θ might be estimated under alternative forms of microdata release (implying alternative specifications of *X*) and a form of release chosen so that θ is inferred to be acceptably small. A sensitivity analysis will usually be required in which the specification of *X* is varied not only according to the form of release but also to allow for alternative plausible forms of external information which an intruder might hold about known population units. For example, one might consider both an intruder with access only to publicly available information, such as the visible characteristics of an individual, and an intruder with access to a private database held by an organisation.

1. C.J. Skinner, University of Southampton, Southampton, United Kingdom, S017 1BJ and R.G. Carter, Statistics Canada, B-2 Jean Talon Building, Ottawa, Ontario, K1A 0T6.

3. Estimation of θ

We suppose that the data consist of the values of X for the sample units. Hence, the sample frequencies f_j are known but the population frequencies F_j are unknown ($j = 1, \ldots, J$). The 'parameter' of interest, θ, is also unknown and must be estimated. We adopt a design-based approach to inference in which the f_j are random and the F_j are fixed. As discussed by SE, the 'parameter', θ, therefore depends on s, unlike standard finite population parameters considered in survey sampling.

SE motivate a point estimator of θ by a resampling argument, which may be generalised to the case of unequal probability sampling, as follows.

Repeat the following steps K times.

Step 1: remove a single unit i from the microdata sample s with probability

$$\alpha_i = \pi_i^{-1} \Big/ \sum_s \pi_i^{-1},$$

where π_i is the (first-order) inclusion probability of unit i;

Step 2: copy the removed unit back into the sample with probability π_i;

Step 3: record whether the removed unit matches a unique record in the microdata and whether this match is correct.

The idea is that Step 1 mimics the intruder's (equal probability) selection of a unit from U (using the inverse sampling idea of Hinkins, Oh and Scheuren 1997). Step 2 mimics the inclusion of that unit in s. The estimator of θ is the empirical proportion of unique matches which are correct. Following the argument of SE, this estimator converges almost surely as $K \to \infty$ to

$$\hat{\theta} = \frac{\sum_{s^{(1)}} \Pr(\text{unit } i \text{ removed and then copied back})}{\left[\begin{array}{l}\sum_{s^{(1)}} \Pr(\text{unit } i \text{ removed and then copied back}) \\ + \sum_{s^{(2)}} \Pr(\text{unit } i \text{ removed and then not copied back})\end{array}\right]}$$

$$= \sum_{s^{(1)}} \alpha_i \pi_i \Big/ \left[\sum_{s^{(1)}} \alpha_i \pi_i + \sum_{s^{(2)}} \alpha_i (1-\pi_i)\right]$$

$$= n^{(1)} \Big/ \left[n^{(1)} + \sum_{s^{(2)}} (\pi_i^{-1} - 1)\right], \qquad (1)$$

where $s^{(1)}$ is the subsample of unique units in s, $s^{(2)}$ is the subsample of units which occur in pairs and $n^{(1)} = \sum_j I(f_j = 1)$ is the size of $s^{(1)}$. In the case of equal probability sampling with $\pi_i = \pi$, $\hat{\theta}$ reduces to $n^{(1)} / [n^{(1)} + 2n^{(2)}(\pi^{-1} - 1)]$, where $2n^{(2)} = 2\sum_j I(f_j = 2)$ is the size of $s^{(2)}$, as in SE.

We are interested in $\hat{\theta}$, defined in (1), as an estimator of θ. SE show that $\hat{\theta}$ is consistent for θ in the equal probability sampling case. The basic steps of their argument may be generalised to the case of unequal probability sampling as follows. We may write

$$\theta = n^{(1)} \Big/ \left[n^{(1)} + \sum_j (F_j - 1) I(f_j = 1) \right]. \qquad (2)$$

Hence, by comparing (1) and (2), $\hat{\theta}$ will be a 'good' estimator of θ if $\sum_{s^{(2)}} (\pi_i^{-1} - 1)$ is a 'good' estimator of $\sum_j (F_j - 1) I(f_j = 1)$. In fact, we prove in Appendix 1 that the latter estimator is unbiased, that is

$$E\left[\sum_{s^{(2)}} (\pi_i^{-1} - 1)\right] = E\left[\sum_j (F_j - 1) I(f_j = 1) \right], \qquad (3)$$

under the assumption of Poisson sampling, that is where population units are sampled independently. Equation (3) generalizes Proposition 2 of SE. In the equal probability sampling case SE show how the result in equation (3) may be extended to prove consistency of $\hat{\theta}$ as an estimator of θ, using an asymptotic framework where $J \to \infty$ and under some regularity conditions, in particular that the F_j are bounded.

Having established the main unbiasedness result in (3), we conjecture that this consistency result will generalise to the case of unequal probability Poisson sampling, subject to additional weak conditions on the π_i, for example that the π_i are bounded above by a positive constant.

The Poisson sampling assumption generalises the Bernoulli sampling assumption in SE. They conclude that in practice $\hat{\theta}$ will remain approximately unbiased for θ under a number of other equal probability sampling designs including simple random sampling, (equal probability) systematic sampling or proportionate stratified simple random sampling. We suggest that in a similar way $\hat{\theta}$ will remain approximately unbiased for θ under corresponding unequal probability designs, *i.e.*, disproportionate stratified simple random sampling and unequal probability systematic sampling. We also suggest that it may be reasonable to allow for nonresponse in $\hat{\theta}$ if s is the set of respondents and if π_i^{-1} consists of a weight which may be interpreted as the reciprocal of the estimated probability of both being sampled and responding.

As discussed in SE, the form of sampling which seems to have the potential to lead to most bias in $\hat{\theta}$ as an estimator of θ in practice is multistage sampling, where the multistage units are strongly related with respect to X. For example, bias might be non-negligible when households form clusters within which all adult individuals are sampled, where the microdata includes individual-level records and where X is primarily determined by household-level variables. This might lead to a higher value of $n^{(2)}/n^{(1)}$ than expected under Poisson sampling and hence to underestimation of θ. Such an example is somewhat contrived, however, and we suspect that the bias of $\hat{\theta}$ as an estimator of θ will be modest in most typical social surveys.

4. Variance Estimation

SE present a linearization estimator of var $(\hat{\theta}-\theta)$, which depends on $n^{(1)}$ and $n^{(2)}$, like $\hat{\theta}$, as well as on $n^{(3)}=\sum_j I(f_j=3)$, the number of values of X for which there are exactly three microdata records. We show in Appendix 2 that this variance estimator may be generalised, in the case of unequal probability Poisson sampling, to

$$\hat{v}=\hat{\theta}^2\frac{\sum_{j=1}^{J}\{I(f_j=3)(\gamma_{1j}^2-\gamma_{2j})+I(f_j=2)(\gamma_{1j}^2-\gamma_{1j})\}}{\left[n^{(1)}+\sum_j I(f_j=2)\,\gamma_{1j}\right]^2} \quad (4)$$

where $\gamma_{1j}=\sum_{s_j}\beta_i$, $\gamma_{2j}=\sum_{s_j}\beta_i^2$, $\beta_i=\pi_i^{-1}-1$ and $s_j=\{i\in s; X_i=j\}$, where X_i is the value of X for unit i.

Note that, in this notation, we may write

$$\hat{\theta}=n^{(1)}\Big/\left[n^{(1)}+\sum_j I(f_j=2)\gamma_{1j}\right].$$

As in the equal probability case, both $\hat{\theta}$ and $\hat{v}$ can be computed straightforwardly from the values X_i and π_i for $i\in s$. The expression given above for $\hat{v}$ reduces to the expression given in Proposition 3 of SE when $\pi_i=\pi$ for all $i\in s$.

The linearisation argument which gives $\hat{v}$ assumes J is large. This seems a weak condition relative to the assumption of Poisson sampling. The linearisation variance estimator does not appear to generalise straightforwardly to other complex sampling designs. This is because the linearised form of $\hat{\theta}-\theta$ depends on the F_j and these cannot simply be replaced by consistent estimators. It also does not appear to be straightforward to apply replication methods to estimate the variance of $\hat{\theta}-\theta$, since θ is unknown and, as indicated by the simulation study in SE, the variance of θ may not be negligible in practice relative to the variance of $\hat{\theta}$.

5. Concluding Remarks

The estimated measure $\hat{\theta}$ considered in this paper may be used as evidence in assessing whether or not a proposed microdata file has an acceptable level of disclosure risk. The aim may be to ensure that $\hat{\theta}$ does not exceed some specified probability. To allow for sampling variation in $\hat{\theta}$ a more conservative procedure would be to require that the upper bound of a confidence interval for θ, say $\hat{\theta}+2\hat{v}^{1/2}$, does not exceed the specified probability.

As well, $\hat{\theta}$ may be used to compare alternative strategies to control disclosure risk. For example, variables may be included in microdata files with more or less classification detail. Greater detail may enhance the value of the file for analysis but may also increase disclosure risk if the variable could be used to match against external information. The estimated measure $\hat{\theta}$ could, therefore, be used to assess the relative risk resulting from different ways of collapsing the level of classification in specific identifying variables, including geography.

The measure may be estimated not only for the population as a whole, but also for subpopulations. Such a breakdown of the measure permits a more realistic assessment of the risk posed by intruders who target specific subpopulations. Such a targeted threat invalidates the basic assumption underlying the definition of whole population measure, θ, that the population unit with known identity is randomly drawn from U with equal probabilities. Separate estimation of the measure in different strata with different sampling fractions also provides a simple method of handling unequal probabilities of selection. This paper has shown how to allow for more general sources of unequal probability sampling in $\hat{\theta}$ and $\hat{v}$. More research is required to assess the robustness of these estimators to departures from Poisson sampling, especially multi-stage sampling.

A potential problem with estimating the measure separately by subpopulations is the impact of the reduction in sample size. SE found $\hat{\theta}$ to be stable in their numerical investigations, with a coefficient of variation never exceeding 6%. Their minimum sample size was, however, about 9,000 so further numerical work is needed to assess the stability of $\hat{\theta}$ for smaller sample sizes. The proposed variance estimation method provides some guidance for any specific case. Stability could, in principle, be improved by the use of model assumptions and one of us (CJS) is conducting further research on the limiting case of a small subpopulation, a single unit, extending θ to a record-level measure of risk analogous to that considered by Skinner and Holmes (1998).

Appendix 1

Proof of Equation (3)

Let $\beta_i=\pi_i^{-1}-1$ and $U_j=\{i\in U; X_i=j\}$, $j=1,\ldots,J$, where X_i denotes the value of X for unit i. The size of U_j is F_j. Instead of labelling units in U by the single index i, consider the double index (jk), $j=1,\ldots,J$, $k=1,\ldots,F_j$, so that, for example, $\pi_{(jk)}$ denotes the inclusion probability for the k^{th} unit in U_j and $\beta_{(jk)}=\pi_{(jk)}^{-1}-1$. Under Poisson sampling the right side of (3) is

$$E\left[\sum_j (F_j-1)I(f_j=1)\right]$$
$$=\sum_{j=1}^{J}(F_j-1)\sum_{k=1}^{F_j}\pi_{(jk)}\prod_{\substack{\ell=1\\ \ell\neq k}}^{F_j}(1-\pi_{(j\ell)}) \quad (A.1)$$

and the left side of (3) is

$$E\left[\sum_{i\in s^{(2)}}\beta_i\right]$$

$$=\sum_{j=1}^{J}\sum_{\substack{k=1\\k<\ell}}^{F_j}\sum_{\ell=1}^{F_j}\pi_{(jk)}\pi_{(j\ell)}\left[\prod_{\substack{m=1\\m\neq k,\,\ell}}^{F_j}(1-\pi_{(jm)})\right]\left[\beta_{(jk)}+\beta_{(j\ell)}\right]$$

$$=\sum_{j=1}^{J}\sum_{\substack{k=1\\k\neq\ell}}^{F_j}\sum_{\ell=1}^{F_j}\pi_{(jk)}\pi_{(j\ell)}\left[\prod_{\substack{m=1\\m\neq k,\,\ell}}^{F_j}(1-\pi_{(jm)})\right]\beta_{(jk)}$$

$$=\sum_{j=1}^{J}\sum_{\substack{k=1\\k\neq\ell}}^{F_j}\sum_{\ell=1}^{F_j}\pi_{(jl)}\left[\prod_{\substack{m=1\\m\neq\ell}}^{F_j}(1-\pi_{(jm)})\right]$$

$$=\sum_{j=1}^{J}(F_j-1)\sum_{\ell=1}^{F_j}\pi_{(j\ell)}\left[\prod_{\substack{m=1\\m\neq\ell}}^{F_j}(1-\pi_{(jm)})\right]$$

which is identical to (A.1) so (3) follows.

Appendix 2

Derivation of Linearisation Variance Estimator

Write $\hat{\theta}-\theta=\tau_1/(\tau_1+\tau_2)-\tau_1/\tau_3$, where

$$\tau_1=\sum_j I(f_j=1),\ \tau_2=\sum_j I(f_j=2)\gamma_{1j},\ \tau_3=\sum_j F_j I(f_j=1).$$

Let $\mu_1=E(\tau_1), t=1,2,3,$ and note that $\mu_1+\mu_2=\mu_3$ from (3). A linearised expression for $\hat{\theta}-\theta$ is $\mu_1(-\tau_1-\tau_2+\tau_3)/\mu_3^2$, the variance of which may be expressed as

$$\begin{aligned}&\mathrm{var}(\hat{\theta}-\theta)\\&\approx \mathrm{var}\left[(\mu_1/\mu_3^2)\sum_{j=1}^{J}\{(F_j-1)I(f_j=1)-\gamma_{1j}I(f_j=2)\}\right]\\&=(\mu_1/\mu_3^2)^2\sum_{j=1}^{J}\left[\begin{array}{l}(F_j-1)^2\Pr(f_j=1)\\ \qquad +E\{\gamma_{1j}^2 I(f_j=2)\}\end{array}\right].\end{aligned}\qquad (A.2)$$

This generalises the expression for the variance in Proposition 3 of SE. The expression for $\hat{v}$ in (4) is obtained by replacing terms in (A.2) by their unbiased estimators. First, μ_1 and μ_3 are estimated by τ_1 and $\tau_1+\tau_2$ respectively so that μ_1/μ_3^2 is estimated by $\hat{\theta}/(\tau_1+\tau_2)$. Next note that

$$E\left[I(f_j=3)(\gamma_{1j}^2-\gamma_{2j})\right]$$

$$=\sum_{k=1}^{F_j}\sum_{\substack{\ell=1\\k\neq\ell\neq m}}^{F_j}\sum_{m=1}^{F_j}\pi_{(jk)}\pi_{(j\ell)}\pi_{(jm)}\left[\prod_{\substack{n=1\\n\neq k,\,\ell,\,m}}^{F_j}(1-\pi_{(jn)})\right]\beta_{(j\ell)}\beta_{(jm)}$$

$$=\sum_{k=1}^{F_j}\sum_{\substack{\ell=1\\k\neq\ell\neq m}}^{F_j}\sum_{m=1}^{F_j}\pi_{(jk)}\left[\prod_{\substack{n=1\\n\neq k}}^{F_j}(1-\pi_{(jn)})\right]$$

$$=(F_j-1)(F_j-2)\Pr(f_j=1),$$

using the notation of Appendix 1. We may also show that

$$E[I(f_j=2)\gamma_{1j}]=(F_j-1)\Pr(f_j=1)\qquad (A.3)$$

by following the proof of (3) in Appendix 1, but omitting the summation over *j*. (Note that the sides of (3) are equal to the corresponding sides of (A.3) summed over *j*). Hence, an unbiased estimator of $(F_j-1)^2\Pr(f_j=1)$ is

$$I(f_j=3)(\gamma_{1j}^2-\gamma_{2j})+I(f_j=2)\gamma_{1j}.$$

It follows that the numerator of the expression for $\hat{v}/\hat{\theta}^2$ in (4) is unbiased for the second part of the expression on the right side of (A.2) (omitting $(\mu_1/\mu_3^2)^2$) as required.

References

Bethlehem, J.G., Keller, W.J. and Pannekoek, J. (1990). Disclosure control of microdata. *Journal of the American Statistical Association*, 85, 38-45.

Hinkins, S., Oh, H.L. and Scheuren, F. (1997). Inverse sampling design algorithms. *Survey Methodology*, 23, 11-21.

Skinner, C.J., and Elliot, M.J. (2002). A measure of disclosure risk for microdata. *Journal of the Royal Statistical Society*, Series B, 64, 855-867.

Skinner, C.J., and Holmes, D.J. (1998). Estimating the re-identification risk per record for microdata. *Journal of Offical Statistics*, 14, 361-372.

Willenborg, L., and De Waal, T. (2001). *Elements of Statistical Disclosure Control*. New York: Springer.