J. R. Statist. Soc. B (2002)
64, *Part* 4, *pp.* 855–867

A measure of disclosure risk for microdata

C. J. Skinner

University of Southampton, UK

and M. J. Elliot

University of Manchester, UK

[Received July 2001. Revised June 2002]

Summary. Protection against disclosure is important for statistical agencies releasing microdata files from sample surveys. Simple measures of disclosure risk can provide useful evidence to support decisions about release. We propose a new measure of disclosure risk: the probability that a unique match between a microdata record and a population unit is correct. We argue that this measure has at least two advantages. First, we suggest that it may be a more realistic measure of risk than two measures that are currently used with census data. Second, we show that consistent inference (in a specified sense) may be made about this measure from sample data without strong modelling assumptions. This is a surprising finding, in its contrast with the properties of the two 'similar' established measures. As a result, this measure has potentially useful applications to sample surveys. In addition to obtaining a simple consistent predictor of the measure, we propose a simple variance estimator and show that it is consistent. We also consider the extension of inference to allow for certain complex sampling schemes. We present a numerical study based on 1991 census data for about 450 000 enumerated individuals in one area of Great Britain. We show that the theoretical results on the properties of the point predictor of the measure of risk and its variance estimator hold to a good approximation for these data.

Keywords: Confidentiality protection; Finite population inference; Sample survey data; Statistical disclosure control

1. Introduction

Anonymized microdata files of individual records from surveys and censuses are often released to researchers so that they may conduct their own analyses. An important consideration for agencies deciding whether and how to release such files is the need to protect against possible statistical disclosure. There is a growing literature on how such protection may take place (Willenborg and de Waal, 2001). A key element of protection methodology is the assessment of disclosure risk for a file. Assessing disclosure risk usually involves difficult judgments (Lambert, 1993) and systematic ways of supporting these judgments by statistical evidence are needed. In this paper we consider one common basic form of such evidence: the values of a single data-based measure of disclosure risk for alternative specifications of the microdata file.

More elaborate forms of evidence, based on statistical modelling frameworks, have been developed to capture more fully the complex nature of potential threats to confidentiality by Paass (1988), Duncan and Lambert (1989), Lambert (1993), Fuller (1993), Skinner *et al*. (1994) and Fienberg *et al*. (1997), among others. We shall not pursue these more general approaches in this paper, however. We restrict attention to three simple measures of disclosure risk. The first

Address for correspondence: C. J. Skinner, Department of Social Statistics, University of Southampton, Highfield, Southampton, SO17 1BJ, UK.
E-mail: cjs@soton.ac.uk

 1369–7412/02/64855

two measures have established uses with 100% census data but suffer difficulties of inference in their extension to sample survey data. The third measure is new and we argue that, not only is it potentially a more realistic measure of risk, but also that it provides a means of overcoming the difficulties of inference for the first two measures. This new measure is the main focus of this paper.

The first measure of risk to be considered is the proportion of units in the population which have unique combinations of values of potentially identifying variables. We denote it Pr(PU), the probability of 'population uniqueness'. It has been used with census microdata in the USA and UK (Greenberg and Voshell, 1990; Marsh *et al*., 1991). Bethlehem *et al*. (1990) discussed Pr(PU), setting out the basic framework which we adopt. The microdata file consists of a set of records for each unit in a sample from a finite population. For a census, only a sample of units is usually included in the file. For a sample survey, the microdata file usually contains records for all units in the sample. Each record contains two disjoint forms of information: identifying information and sensitive information. The former consists of the values of a set of *identifying variables*, which might be matchable to known units in the population. The threat of disclosure arises from the possibility that an *intruder* might succeed in identifying a microdata unit through such matching and hence be able to disclose the sensitive information on this unit. The identifying variables are assumed to be categorical, which is a realistic assumption in many censuses and social surveys. Population uniqueness is considered further by Greenberg and Zayatz (1992) and Skinner *et al*. (1994).

Since only records which are sample unique can be population unique, an alternative measure of risk is the proportion of sample unique records which are population unique, denoted Pr(PU|SU). This is our second measure. It has been used with census microdata in Canada and the UK (Carter *et al*., 1991; Elliot *et al*., 1998; Dale and Elliot, 2001) and has been considered further by Skinner *et al*. (1994), Chen and Keller-McNulty (1998), Samuels (1998) and Fienberg and Makov (1998). The third measure of risk again refers to the threat that is represented by the sample unique records but now allows for the risk arising from records which are not population unique. It is defined as the proportion of correct matches among those population units which match a sample unique microdata record and is denoted θ. The basic idea was introduced by Elliot (2000). We discuss these three measures further in Section 2.

We claim two advantages for the third measure. First, we suggest that it may be a more realistic measure of disclosure risk. Second, we show in Section 3 that consistent inference (in a specified sense) about θ may be made simply, without strong modelling assumptions. This is a surprising finding in its contrast with the property discussed in Section 2, that consistent inference about the 'similar' measures Pr(PU) or Pr(PU|SU) is problematic in the absence of strong modelling assumptions. For a census, these measures may be calculated from the population data, even if a microdata file is released only for a sample. For a sample survey, it is necessary to make inference about the measures from sample data.

The results in Section 3 assume Bernoulli sampling. The extension of these results to other designs is considered in Section 4. A numerical study of the properties of the inference procedures of Section 3 is presented in Section 5. Some concluding remarks are made in Section 6.

2. Measures of disclosure risk

2.1. Framework and notation

Let the microdata file consist of a set of records, each corresponding to a unit in a *microdata sample s*, selected from a finite population U ($s \subset U$). Let n and N denote the numbers of units in s and U respectively. We assume that each record contains the values for the unit on the categor-

ical identifying variables, supposed given. The categorical variable formed by cross-classifying all the identifying variables is denoted X, with values denoted $1, \ldots, J$. Each of these values corresponds to a possible combination of values of the identifying variables. For example, if these variables are sex, age, occupation and marital status then a possible value of X might be (female, 38 years, medical professional, divorced). In practice we may expect J, the number of categories of X, to be large. Let X_i denote the value of X for unit i ($i \in U$). Let the *population frequencies* for the different values of X be denoted

$$F_j = \sum_{i \in U} I(X_i = j), \qquad j = 1, \ldots, J,$$

where $I(\cdot)$ is the indicator function: $I(A) = 1$ if A is true and $I(A) = 0$ otherwise. Categories with zero counts are excluded so that $F_j \geqslant 1$ for $j = 1, \ldots, J$. Let the *population frequencies of frequencies* be denoted

$$N_r = \sum_{j=1}^{J} I(F_j = r), \qquad r = 1, 2, \ldots.$$

For example, N_1 is the number of values of X which are unique in the population. We refer to such a value of X (with $F_j = 1$) as *population unique*. We also describe a unit as population unique if its value is population unique. Note that

$$\begin{aligned} \sum_{r=1}^{\infty} N_r &= J, \\ \sum_{r=1}^{\infty} r N_r &= N. \end{aligned} \tag{1}$$

The sample quantities f_j and n_r are defined analogously to F_j and N_r respectively. Thus, the *sample frequency* for value j of X is denoted

$$f_j = \sum_{i \in s} I(X_i = j), \qquad j = 1, \ldots, J,$$

and the *sample frequencies of frequencies* are denoted

$$n_r = \sum_{j=1}^{J} I(f_j = r), \qquad r = 0, 1, 2\ldots.$$

A value j of X is called *sample unique* if $f_j = 1$. Similarly, a unit is called sample unique if its value of X is sample unique.

2.2. Three measures of disclosure risk

The first measure to be considered is $N_1/N = \Sigma I(F_j = 1)/N$, the proportion of units in the population which are population unique. To interpret N_1/N as a measure of risk, note first that any population unique unit might be viewed as 'risky'. For, if an intruder could use X to link an identifiable unit in the population to a unit in the microdata sample and know that the unit was population unique then the intruder would know that the link was correct (provided that X was measured in the same way). The measure N_1/N is the probability that a unit is population unique under the 'method of attack' in which the intruder draws the unit at random (with equal probabilities) from the population. We write $N_1/N = \Pr(\mathrm{PU})$.

The second measure to be considered restricts attention to sample unique microdata units, since only these can be population unique, and is defined as the probability that a sample unique

unit is population unique, under the method of attack in which the intruder draws the unit at random (with equal probabilities) from the sample unique units:

$$\Pr(\mathrm{PU}|\mathrm{SU}) = \sum_j I(f_j = 1, F_j = 1) \Big/ \sum_j I(f_j = 1).$$

Finally, the proposed measure of disclosure risk is given by

$$\theta = \sum_j I(f_j = 1) \Big/ \sum_j F_j \, I(f_j = 1). \tag{2}$$

To interpret θ, consider the method of attack in which an intruder draws a unit randomly (with equal probabilities) from the population. Call this the *chosen unit*. The intruder then matches the value of X for the chosen unit to the value of X for each unit in the microdata sample s. A *unique match* is said to be established if there is just one unit in s with the same value of X. Call this the *matching unit*. A unique match is said to be a *correct match* if the matching unit and the chosen unit are identical. The number of possible chosen units for which a unique match will exist is $\Sigma F_j \, I(f_j = 1)$, the denominator of θ, and the number of these units for which the match is correct is $\Sigma I(f_j = 1)$, the numerator of θ. Hence we may write $\theta = \Pr(\text{correct match}|\text{unique match})$ and interpret θ as the conditional probability that a unique match will be correct under this method of attack, which we call *attack method 1*.

The definitions of both Pr(PU|SU) and θ depend via the f_j on the sample s and these measures are thus not conventional finite population parameters of the kind that is considered in survey sampling (Cochran, 1977). That a measure of disclosure risk should be sample dependent is natural, since it is only the records that are released in the sample that are at risk of identification (see Lambert (1993), page 322). Because Pr(PU|SU) and θ are sample dependent, we shall refer to inference about these measures as prediction rather than estimation.

2.3. Discussion and comparison of three measures

A problem with the interpretation of Pr(PU) as a measure of risk for all microdata records is that not all sample units are 'equally likely' to be population unique. In particular, if a record is not sample unique then it cannot be population unique. Thus the proportion of population unique units among all microdata records (which approximates Pr(PU) under sampling with equal inclusion probabilities) will not exceed Pr(PU|SU), the proportion of population unique units among sample unique microdata records. The measure Pr(PU) may therefore be rejected in favour of Pr(PU|SU) on the grounds that Pr(PU) is too optimistic a measure. It is possible to extend this argument to argue that not all sample unique microdata records are equally likely to be population unique (Skinner and Holmes, 1998). Such extensions involve complications in modelling, however, which we wish to avoid. We restrict attention to measures which take a single value for the microdata file.

To compare Pr(PU|SU) and θ as measures of risk, we extend the method of attack under which Pr(PU|SU) is defined.

In *attack method 2* the intruder draws one microdata record at random (with equal probabilities) from the sample unique records and searches through the population at random until a unit is located which matches the selected record.

Under this method, the probability P that the selected record belongs to the unit located is at least Pr(PU|SU), since

$$P = \sum_j I(f_j = 1) F_j^{-1} \Big/ \sum_j I(f_j = 1) \tag{3}$$

$$\geqslant \sum_j I(f_j = 1, F_j = 1) \Big/ \sum_j I(f_j = 1) = \Pr(\text{PU}|\text{SU}).$$

This inequality suggests that Pr(PU|SU) is an overoptimistic measure since it fails to reflect the risk arising from values of X which are twins ($F_j = 2$), triples ($F_j = 3$) and so forth. Like θ, P may be interpreted as Pr(correct match|unique match), but under attack method 2 rather than under attack method 1. Whether θ or P is a more appropriate measure of risk depends on the method of attack. We suggest that method 1 is more plausible than method 2 since the intruder makes fuller use of all the microdata information in method 1. For example, this method is similar to that employed by Blien *et al*. (1992), who matched all records in the microdata against all records in an external file. We therefore claim that θ is a useful measure of disclosure risk.

2.4. Inference about Pr(PU) and Pr(PU|SU)

Inference about Pr(PU) or Pr(PU|SU) appears to be an intrinsically difficult problem. Assuming that N is known, the estimation of Pr(PU) reduces to the estimation of N_1, which appears to share similar problems to the well-known difficulties that are involved in estimating $J = \Sigma r N_r$ (Bunge and Fitzpatrick, 1993). A natural approach is to write

$$E(n_r) = \sum_s N_s P_{rs}, \qquad r = 1, 2, \ldots, \tag{4}$$

where $E(\cdot)$ is the expectation with respect to sampling and the coefficients P_{rs} are known for sampling schemes such as simple random sampling or Bernoulli sampling (Goodman, 1949). The solution of these equations for N_r with $E(n_r)$ replaced by n_r gives unbiased estimators of J and N_1 under apparently weak conditions (Goodman, 1949). Unfortunately, Goodman found that the estimator of J can be 'very unreasonable' and the same appears to be so for the corresponding estimator of N_1 (given in his theorem 4). One interpretation is that this is a problem of collinearity between the equations in expression (4). An alternative 'nonparametric' estimator of N_1 has been proposed by Zayatz (1991) and Greenberg and Zayatz (1992) but appears to be subject to serious upward bias for small sampling fractions (Chen and Keller-McNulty, 1998).

One way of addressing the estimation difficulties is to make stronger modelling assumptions. Bethlehem *et al*. (1990) set out one approach based on the Poisson–gamma model but this approach appears not to be robust, as discussed by Skinner *et al*. (1994) and Chen and Keller-McNulty (1998). The latter proposed an estimator based on a slide negative binomial model which improved on the Poisson–gamma model but still had upward bias for small sampling fractions when J is known and was found to be unstable for small sampling fractions when J is unknown (the usual case).

Samuels (1998) discussed the point prediction of Pr(PU|SU) based on a Poisson–Dirichlet model. Although he obtained some encouraging results, he found substantial underprediction when the sampling fraction is low and commented in his section 6 on the intrinsic difficulties in making inference about Pr(PU|SU) in certain situations. In summary, we suggest that no inference procedure is currently available which robustly estimates Pr(PU) or predicts Pr(PU|SU) across the wide range of possible population structures that may exist in surveys and for small sampling fractions. In the next section we show how θ may be predicted without strong modelling assumptions.

3. Inference about the measure proposed

In this section we consider inference about θ in equation (2). We assume that the sample frequencies of frequencies n_r, $r = 1, 2, \ldots,$ are known but that the F_j and N_r are unknown. We adopt a design-based survey sampling framework in which the finite population quantities F_j and N_r are fixed and the only source of randomness comes in the selection of the sample s. As a consequence, not only are the sample quantities f_j and n_r random but so also is the 'parameter' θ of interest (see the discussion in Section 2.2) and so we refer to prediction of θ rather than estimation. For simplicity, we shall assume Bernoulli sampling in which all population units are sampled independently with a common probability π. We consider extensions to other sampling designs in Section 5. In particular, the Bernoulli sampling assumption implies that the f_j are independently binomially distributed:

$$f_j \sim \text{bin}(F_j, \pi), \qquad j = 1, \ldots, J. \tag{5}$$

To motivate our point predictor of θ, we consider a simulation-based approach, which mimics attack method 1, referred to here as *data intrusion simulation*. Repeat the following steps (independently) for $k = 1, 2, \ldots, K$.

Step 1: remove one unit at random (with equal probabilities) from the sample.
Step 2: copy the unit back into the sample with probability π.
Step 3: record whether the unit removed has a unique match on X with a sample unit ($R_{uk} = 1$ if so; $R_{uk} = 0$ otherwise) and, if so, whether this match is correct ($R_{ck} = 1$ if so; $R_{ck} = 0$ otherwise).

The predictor of θ is then the proportion of unique matches which are correct:

$$\hat{\theta}(K) = \sum_{k=1}^{K} R_{ck} R_{uk} \Big/ \sum_{k=1}^{K} R_{uk}. \tag{6}$$

Step 1 simulates a random draw from the population, as in attack method 1, since the sample units are assumed to be drawn with equal probabilities. Likewise, step 2 simulates the fact that the population unit that is selected by the intruder will be included in the sample with probability π. The predictor $\hat{\theta}(K)$ is formed from the usual 'analogue' principle which 'estimates' a population quantity by the corresponding sample quantity. This principle does not generate sensible inference about Pr(PU) or Pr(PU|SU), however, and so it is natural to be sceptical about whether $\hat{\theta}(K)$ will be a sensible predictor of θ. Having used the data intrusion simulation and the analogue principle to motivate the form of $\hat{\theta}(K)$, we note that the limit of $\hat{\theta}(K)$ as $K \to \infty$ can, in fact, be expressed simply in closed form.

Proposition 1. $\hat{\theta}(K) \to \hat{\theta}$ almost surely (with respect to the randomization in the simulation), where

$$\hat{\theta} = \frac{\pi n_1}{\pi n_1 + 2(1 - \pi) n_2}. \tag{7}$$

The proofs of this and subsequent propositions are given in Appendix A.

Since n_1 and n_2 are easy to compute and π is known, $\hat{\theta}$ is also easy to compute and, we propose, is used in practice rather than $\hat{\theta}(K)$. The latter predictor has been introduced only to motivate $\hat{\theta}$ and possible extensions.

We now assess whether $\hat{\theta}$ in equation (7) is a sensible predictor of θ in equation (2). To consider this, $\hat{\theta}$ and θ may be re-expressed as

$$\hat{\theta} = \frac{n_1}{n_1 + 2(1-\pi)n_2/\pi},$$

$$\theta = \frac{n_1}{n_1 + \sum_j (F_j - 1)\, I(f_j = 1)}.$$

Hence $\hat{\theta}$ will be a sensible predictor of θ if n_2 is a sensible predictor of

$$\frac{\pi \sum (F_j - 1)\, I(f_j = 1)}{2(1-\pi)}.$$

The unbiasedness of the latter predictor with respect to the Bernoulli sampling is implied by the following result.

Proposition 2.

$$E\{2(1-\pi)n_2/\pi\} = E\Big\{ \sum_j (F_j - 1)\, I(f_j = 1) \Big\}. \tag{8}$$

This proposition provides an alternative expression for one of the set of equations in expression (4) and shows that prediction of θ by $\hat{\theta}$ may be obtained by solving only one of these equations (with $E(n_r)$ replaced by n_r as in Goodman (1949)) rather than the entire set as required for N_1 or J.

To study the properties of $\hat{\theta}$ as a predictor of θ further, we consider an asymptotic framework involving a sequence of populations $U^{(t)}$ of sizes $N^{(t)}$, with $N^{(1)} < N^{(2)} < \ldots$, indexed by $t = 1, 2, \ldots$. We suppose that the specification of the identifying variables also depends on t and we write $X^{(t)}$, with corresponding frequencies $F_j^{(t)}$ in $U^{(t)}$ for $J^{(t)}$ categories j (for which $F_j^{(t)} \geqslant 1$). We assume that $0 < J^{(1)} < J^{(2)} < \ldots$, reflecting the idea that an intruder will need to use identifying variables with more detail to identify units in larger populations. In practical applications of the methods of this paper it is assumed that J will be large; for example J is around 4×10^6 in the application in Section 6. The sample frequencies are denoted $f_j^{(t)}$, $j = 1, \ldots, J^{(t)}$, and it is assumed, as in distribution (5), that for each t the $f_j^{(t)}$ are independently distributed as $\text{bin}(F_j^{(t)}, \pi)$, where π does not depend on t. The sequence of random variables $f_1^{(t)}, \ldots, f_{J^{(t)}}^{(t)}$ defines a double array (Serfling (1980), page 31) for $t = 1, 2, \ldots$. No assumption is made about the nesting of the populations for $t = 1, 2, \ldots$. The key assumption used to apply results for double arrays to obtain the limiting properties of $\hat{\theta}^{(t)} - \theta^{(t)}$ is that the $F_j^{(t)}$ are bounded.

Assumption 1. There is a constant $C < \infty$ such that $1 \leqslant F_j^{(t)} < C$ for $j = 1, \ldots, J^{(t)}, t = 1, 2, \ldots$. In our asymptotic framework the sample size $n^{(t)} = \Sigma f_j^{(t)}$ will converge to ∞ almost surely, along with the population size $N^{(t)}$ (assuming that $\pi > 0$). Henceforth, we shall suppress the superscript t for simplicity, other than writing $t \to \infty$ to represent limits in this asymptotic framework.

Proposition 3. Under Bernoulli sampling with sampling fraction π, $0 < \pi < 1$, and assumption 1, we have $\hat{\theta} - \theta = o_p(1)$ and $(\hat{\theta} - \theta)/v^{1/2} \to N(0, 1)$ as $t \to \infty$, where

$$v = c^2 \sum_{j=1}^{J} F_j(F_j - 1)(\pi F_j - 3\pi + 2)(1-\pi)^{F_j - 1},$$

$$c = \sum F_j \pi (1-\pi)^{F_j - 1} \Big/ \Big\{ \sum F_j^2 \pi (1-\pi)^{F_j - 1} \Big\}^2.$$

Thus, under the given asymptotic framework, $\hat{\theta}$ is a consistent predictor of θ in the sense that $\hat{\theta} - \theta$ converges in probability to 0. A simple estimator of the variance of $\hat{\theta} - \theta$ is

$$\hat{v} = \hat{\theta}^2 \frac{2(1-\pi)\{3(1-\pi)n_3 + (2-\pi)n_2\}}{\{\pi n_1 + 2(1-\pi)n_2\}^2}. \tag{9}$$

This estimator is consistent for v in the following sense.

Proposition 4. Under the assumptions of proposition 3, $\hat{v} = v + o_p(J^{-1})$.

A corollary of propositions 3 and 4 is, from Slutsky's lemma, that

$$(\hat{\theta} - \theta)/\hat{v}^{1/2} \to N(0, 1) \qquad \text{as } t \to \infty.$$

For the disclosure control application, an agency might adopt a conservative approach by considering the upper bound of a one-sided prediction interval $\hat{\theta} + z_{1-\alpha}\hat{v}^{1/2}$, where $z_{1-\alpha}$ is the $(1-\alpha)$-quantile of the standard normal distribution, and requiring this bound to be sufficiently low, say below 0.25.

4. Complex sampling designs

The results so far have assumed Bernoulli sampling for simplicity. In this section we consider the extension to other survey sampling designs.

4.1. Simple random sampling without replacement

Under simple random sampling of size n, the binomial distribution of f_j in distribution (5) is replaced by the hypergeometric distribution with parameters (N, n, F_j). We define $\hat{\theta}$ as in equation (7) with $\pi = n/N$. It may again be shown that $\hat{\theta}$ is consistent for this design, in the sense that $\hat{\theta} - \theta = o_p(1)$, under the same asymptotic framework as in Section 3, where now the sample size $n^{(t)}$ is determined by the design and it is supposed that, as $t \to \infty$, $n^{(t)} \to \infty$ and $n^{(t)}/N^{(t)} \to \pi_0$, a non-zero constant. This consistency property may be proved by showing that the numerators and denominators of $\hat{\theta}$ and θ converge in mean square to the same non-zero limits as $t \to \infty$, assuming again that the $F_j^{(t)}$ are bounded (assumption 1) and that the sequences $(F_1^{(t)}, \ldots, F_{J^{(t)}}^{(t)})$ are such that these limits exist. The key additional complication is that the f_j are no longer independent. The correlations between the f_j, arising from their multiple hypergeometric distribution, tend to 0, however, as $t \to \infty$ and do not prevent the consistency of $\hat{\theta}$.

4.2. Proportionate stratification

In the case of stratified sampling, the population consists of H strata of sizes $N_1, \ldots, N_H$ and independent simple random samples of sizes $n_1, \ldots, n_H$ are drawn from these strata. Letting f_{hj} and F_{hj} be defined analogously to f_j and F_j within stratum h (so that $\Sigma_h f_{hj} = f_j$ and $\Sigma_h F_{hj} = F_j$), the distribution of f_{hj} is now hypergeometric with parameters (N_h, n_h, F_{hj}). If $n_h/N_h = \pi$ then $\hat{\theta}$ may be defined as in equation (7). If we assume an asymptotic framework in which $n_h \to \infty$, $N_h \to \infty$, $n_h/N_h \to \pi$ (fixed) and F_j is bounded ($h = 1, \ldots, H$; $j = 1, \ldots, J$) then it may be argued that $\hat{\theta} = \theta + o_p(1)$, as in Section 4.1.

4.3. Unequal probability sampling

More generally, the inclusion probabilities π_i of different population units may be unequal. The simplest case is Poisson sampling when different units are selected independently. The definition of θ remains unchanged, since Pr(correct match|unique match) is not defined with respect to

the sampling mechanism. It is not possible, however, to maintain the definition of $\hat{\theta}$ in equation (7), since it depends on π, an assumed common inclusion probability. The modification of $\hat{\theta}$ to achieve consistent prediction of θ under general unequal probability designs is thus a topic for further research.

4.4. Cluster sampling

Under cluster sampling, it is possible for $\hat{\theta}$ in equation (7) to be seriously inconsistent for θ, even if all units have a common inclusion probability π.

Suppose that the population is partitioned into clusters of size 1 or 2 and cluster sampling is employed with equal inclusion probability π. Suppose that X takes a common value for two units in the same cluster but different values for two units in different clusters. Hence $F_j = 1$ or $F_j = 2$ and $f_j = 0$ or $f_j = F_j$ for all j. It follows from equation (2) that $\theta = 1$. But n_2 will not in general be 0 and the probability limit of $\hat{\theta}$ will in general be less than 1 and may be arbitrarily close to 0, dependent on π and the proportion of clusters of size 2 in the population. It is clear that $\hat{\theta}$ may be a poor predictor of θ in this example.

To obtain a heuristic guide to the effect of cluster sampling, note that for $\hat{\theta}$ to be a reasonable predictor of θ the ratios $\Pr(f_j = 2)/\Pr(f_j = 1)$ should roughly follow that for the binomial distribution of expression (5) (so that proposition 2 holds approximately). If the clusters are defined in a way that is fairly unrelated to X then this condition may hold and it seems plausible that this will be so for many social surveys. The predictor $\hat{\theta}$ will be most harmed when the cluster sampling has a clear direct effect on $\Pr(f_j = 1)$ or $\Pr(f_j = 2)$. Consider, for example, a survey of adults in which households form clusters (all adults in the household are sampled) and the variables defining X are all defined at the household level, even though θ is defined at the individual level. This form of sampling may tend to distort the ratio $\Pr(f_j = 2)/\Pr(f_j = 1)$ relatively to what would be expected under Bernoulli sampling of individuals.

5. Numerical study

The aim of this section is to provide some numerical evidence on the sampling properties of θ, $\hat{\theta}$ and $\hat{v}$, defined in Sections 2 and 3. We do not consider Pr(PU) or Pr(PU|SU) because of the absence of suitable inference procedures, as discussed in Section 2.4. We use data from the 1991 population census of Great Britain on all enumerated individuals in one area of around $N = 450\,000$ people. The variable X was formed by cross-classifying the following potential identifying variables (with numbers of categories in parentheses): age group (94), sex (2), marital status (5), ethnic group (10), primary economic status (11) and country of birth (42). This choice of identifying variables was based on the discussion of possible scenarios of attack by an intruder in Elliot and Dale (1999). Many different cross-classifications have also been investigated and have yielded results that are similar to those presented here.

Samples were drawn from this population by using systematic sampling of 1 in L units for $L = 10, 20, 50$, i.e. with $\pi = 0.1, 0.05, 0.02$, within geographical strata. The stratification follows that used to draw the individual sample of anonymized records, a microdata file released from the 1991 census (Marsh, 1993). Hence, this study provides evidence on the extent to which users of such microdata could infer the value of θ under different sampling fractions (the sampling fraction used in 1991 for individual microdata was 0.02). Within the strata, the individuals in the population are ordered by geography for the systematic sampling and, in this way, further implicit stratification by geography is achieved. By departing from the Bernoulli sampling assumption, this study provides some evidence on the robustness of the results of Section 3 to

alternative sampling assumptions. A further advantage of the use of systematic sampling is that we may evaluate the exact bias and variance of $\hat{\theta}$ and its standard error estimator $\hat{v}^{1/2}$ by enumerating all L possible samples.

Table 1 contains the means and standard deviations across the L systematic samples for θ, $\hat{\theta}$ and $\hat{v}^{1/2}$ defined in equations (2), (7) and (9) respectively. Considering the measure θ first, we recall that it is not a fixed population parameter but is sample dependent. As expected, θ tends to decrease as π decreases, reflecting the disclosure protection of sampling. For a fixed sample size, however, the results do not indicate great sampling variation in θ. For example, for a 2% sampling fraction, θ varies only between 4.1% and 4.6% across the 50 possible systematic samples. Turning to the predictor $\hat{\theta}$, we define its bias by the mean of $\hat{\theta} - \theta$. We see that for each sampling fraction the (absolute) bias is smaller than 16% of the standard error and is never greater than 0.1% in absolute terms. This seems likely to be acceptably small for most practical applications. Furthermore, the standard error of $\hat{\theta}$ (sd$(\hat{\theta} - \theta)$) is also small relative to the mean of $\hat{\theta}$. The coefficient of variation of $\hat{\theta}$ is 5.8%, 4.8% and 3.1% for $\pi = 0.02$, $\pi = 0.05$ and $\pi = 0.10$ respectively so $\hat{\theta}$ is a fairly stable predictor of θ here. The estimator $\hat{v}^{1/2}$ of the standard error of $\hat{\theta}$ does appear to be approximately unbiased. There is a slight upward bias (implying that $\hat{v}$ is a conservative variance estimator) arising perhaps from the stratified systematic design reducing the variance of $\hat{\theta} - \theta$. The coefficient of variation of $\hat{v}^{1/2}$ is 7.6%, 8.7% and 3.2% for $\pi = 0.02$, $\pi = 0.05$ and $\pi = 0.10$ respectively and so $\hat{v}$ is also a fairly stable estimator here. Normal Q–Q-plots of the L values of $\hat{\theta} - \theta$ and of $(\hat{\theta} - \theta)/\hat{v}^{1/2}$ for each sampling fraction suggest that the normal distribution provides a good approximation to the distribution of these errors.

6. Concluding remarks

The measure proposed might be used by a statistical agency choosing between alternative ways of releasing microdata from a sample survey. For example, the agency may consider more or less detailed classifications of potential identifying variables, such as occupation. The value of $\hat{\theta}$ could be calculated for each alternative form of release. The upper bound of a one-sided confidence interval for θ (say $\hat{\theta} + 2.3\hat{v}^{1/2}$ for a 99% interval) might also be computed. Disclosure risk may be assessed either in a relative way, by comparing alternative release strategies, or in an

Table 1. Means and standard deviations of various quantities across all possible systematic samples of a specified sampling fraction from a census population of 450 000 individuals

Parameter		*Results for the following sampling fractions π:*		
		0.02	*0.05*	*0.10*
Risk measure θ	Mean	0.0426	0.1047	0.1985
	Standard deviation	0.0012	0.0051	0.0027
Estimator $\hat{\theta}$	Mean	0.0429	0.1055	0.1990
	Standard deviation	0.0020	0.0058	0.0045
Error $\hat{\theta} - \theta$	Mean	0.0004	0.0008	0.0005
	Standard deviation	0.0025	0.0051	0.0061
Standard error estimator, $\hat{v}^{1/2}$	Mean	0.0028	0.0052	0.0072
	Standard deviation	0.0002	0.0004	0.0002

absolute way, e.g. by requiring that $\hat{\theta}$ (or $\hat{\theta} + 2.3\hat{v}^{1/2}$) may not exceed some specified probability such as 0.25.

We have shown theoretically that our approach can accommodate Bernoulli, simple random or proportionate stratified sampling and have shown numerically that it can accommodate stratified systematic sampling. The extension of our approach to unequal probability sampling and multistage sampling requires further research. Nevertheless, our approach may be applied within strata when stratum sampling fractions are unequal and we conjecture that our approach will be reasonably robust under a self-weighting multistage design, where the multistage units are not strongly related to the categories of X defined by potential identifying variables.

A further advantage of the proposed measure θ is that its definition of Pr(correct match | unique match) may be extended naturally to allow for X to be measured differently in the microdata and in the external information that is available to the intruder, i.e. to allow for measurement error (Fuller, 1993; Fienberg *et al.*, 1997). The data intrusion simulation of Section 3 may be analogously extended under a given misclassification mechanism. The properties of the resulting estimator of θ require further research.

Appendix A: Proofs

A.1. Proof of proposition 1

Observe first that $R_{uk} = 0$ at iteration k unless either of the following two events occurs: *event 1*, a sample unique unit ($f_j = 1$) is drawn at step 1 and is copied back at step 2, so that $R_{uk} = 1$ and $R_{ck} = 1$; *event 2*, a sample twin ($f_j = 2$) is drawn at step 1 and is not copied back at step 2, so that $R_{uk} = 1$ and $R_{ck} = 0$. Hence at each iteration

$$\begin{aligned} \Pr(R_{uk} = 1) &= \{\pi n_1 + 2(1-\pi)n_2\}/n, \\ \Pr(R_{ck} = 1 | R_{uk} = 1) &= \hat{\theta}. \end{aligned} \tag{10}$$

Since the pairs (R_{ck}, R_{uk}) are independent and identically distributed, the proposition follows from the strong law of large numbers, provided that $\Pr(R_{uk} = 1)$ is non-zero.

A.2. Proof of proposition 2

It follows from distribution (5) that both sides of equation (8) equal

$$\sum_j F_j(F_j - 1)\pi(1-\pi)^{F_j - 1}.$$

A.3. Outline proof of proposition 3

We may write $\hat{\theta} - \theta = g(T_J)$, where $T_J = \Sigma Y_j$ and $Y_j = (I(f_j = 1), I(f_j = 2), F_j\, I(f_j = 1))'$. The assumptions of the proposition are sufficient for a central limit theorem (see Serfling (1980), page 32) for the independent random vectors $Y_1, \ldots, Y_J$, giving $V_J^{-1/2}(T_J - \mu_J) \to N(0, I)$ as $t \to \infty$ where $\mu_J = E(T_J)$ and $V_J = \mathrm{var}(T_J)$ are defined with respect to the binomial distributions in expression (5). It follows by the delta method that

$$\{\hat{\theta} - \theta - g(\mu_J)\}/v^{1/2} \to N(0, 1)$$

where $v = \mathrm{var}\{\nabla'(T_J - \mu_J)\}$ and $\nabla = g'(\mu_J)$ is the vector of derivatives of $g(T_J)$ evaluated at $T_J = \mu_J$. Writing $\mu_J = (\mu_{J1}, \mu_{J2}, \mu_{J3})'$ and $\phi = 2(1-\pi)/\pi$, note first that

$$g(\mu_J) = \frac{\mu_{J1}}{\mu_{J1} + \phi\mu_{J2}} - \frac{\mu_{J1}}{\mu_{J3}} = 0$$

since, from proposition 2, $\mu_{J3} = \mu_{J1} + \phi\mu_{J2}$. Next, note that

$$\begin{aligned}\nabla' &= \left(\frac{\phi\mu_{J2}}{(\mu_{J1} + \phi\mu_{J2})^2} - \frac{1}{\mu_{J3}}, -\frac{\phi\mu_{J1}}{(\mu_{J1} + \phi\mu_{J2})^2}, \frac{\mu_{J1}}{\mu_{J3}^2} \right) \\ &= \mu_{J1}\mu_{J3}^{-2}\,[-1, -\phi, 1]\,,\end{aligned}$$

using again the fact that $\mu_{J3} = \mu_{J1} + \phi\mu_{J2}$. Hence

$$v = (\mu_{J1}\mu_{J3}^{-2})^2 \operatorname{var}(-T_{J1} - \phi T_{J2} + T_{J3})\,, \tag{11}$$

where $T_J = (T_{J1}, T_{J2}, T_{J3})'$. Now

$$-T_{J1} - \phi T_{J2} + T_{J3} = \sum_j (F_j - 1)\, I(f_j = 1) - \phi n_2.$$

Using proposition 2, we have

$$\begin{aligned}\operatorname{var}(-T_{J1} - \phi T_{J2} + T_{J3}) &= \sum_j E\{(F_j - 1)\, I(f_j = 1) - \phi\, I(f_j = 2)\}^2 \\ &= \sum_j E\{(F_j - 1)^2\, I(f_j = 1)\} + \phi^2\, \Pr(f_j = 2) \\ &= \sum_j (F_j - 1)^2 F_j \pi(1-\pi)^{F_j - 1} + \phi^2 F_j (F_j - 1)\pi^2 (1-\pi)^{F_j - 2}/2 \\ &= \sum_j F_j (F_j - 1)(1-\pi)^{F_j - 1}\{(F_j - 1)\pi + 2(1-\pi)\} \\ &= \sum_j F_j (F_j - 1)(\pi F_j - 3\pi + 2)(1-\pi)^{F_j - 1}. \end{aligned} \tag{12}$$

Finally the expression for v in proposition 3 is obtained from equations (11) and (12) by noting that

$$\begin{aligned}\mu_{j1} &= \sum_j F_j \pi (1-\pi)^{F_j - 1}, \\ \mu_{j3} &= \sum_j F_j^2 \pi (1-\pi)^{F_j - 1}.\end{aligned}$$

A.4. Outline proof of proposition 4

Note that $\hat{\theta} = \mu_{J1}/\mu_{J3} + o_p(1)$, $\{\pi n_1 + 2(1-\pi)n_2\}/J = \pi\mu_{J3}/J + o_P(1)$ and

$$\frac{3(1-\pi)n_3 + (2-\pi)n_2}{J} = \frac{\left\{\sum F_j(F_j - 1)(\pi F_j - 3\pi + 2)(1-\pi)^{F_j - 1}\right\}\pi^2}{2J(1-\pi)} + o_p(1).$$

It follows that

$$\hat{v} = \mu_{J1}^2 \mu_{J3}^{-4} \left\{\sum F_j(F_j - 1)(\pi F_j - 3\pi + 2)(1-\pi)^{F_j - 1}\right\} + o_p(J^{-1})$$

and the result follows since $c = \mu_{J1}/\mu_{J3}^2$.

References

Bethlehem, J. G., Keller, W. J. and Pannekoek, J. (1990) Disclosure control of microdata. *J. Am. Statist. Ass.*, **85**, 38–45.

Blien, U., Wirth, H. and Müller, M. (1992) Disclosure risk for microdata stemming from official statistics. *Statist. Neerland.*, **46**, 69–82.

Bunge, J. and Fitzpatrick, M. (1993) Estimating the number of species: a review. *J. Am. Statist. Ass.*, **88**, 364–373.

Carter, R., Boudreau, J.-R. and Briggs, M. (1991) Analysis of the risk of disclosure for census microdata. *Report*. Social Survey Methods Division, Statistics Canada, Ottawa.

Chen, G. and Keller-McNulty, S. (1998) Estimation of identification disclosure risk in microdata. *J. Off. Statist.*, **14**, 79–95.

Cochran, W. G. (1977) *Sampling Techniques*, 3rd edn. New York: Wiley.

Dale, A. and Elliot, M. (2001) Proposals for 2001 samples of anonymized records: an assessment of disclosure risk. *J. R. Statist. Soc.* A, **164**, 427–447.
Duncan, G. T. and Lambert, D. (1989) The risk of disclosure for microdata. *J. Bus. Econ. Statist.*, **7**, 207–217.
Elliot, M. J. (2000) DIS: a new approach to the measurement of statistical disclosure risk. *Int. J. Risk Mangmnt*, **2**, 39–48.
Elliot, M. J. and Dale, A. (1999) Scenarios of attack: the data intruder's perspective on statistical disclosure risk. *Nethrlnds Off. Statist.*, spring, 6–10.
Elliot, M. J., Skinner, C. J. and Dale, A. (1998) Special uniques, random uniques and sticky populations: some counterintuitive effects of geographical detail on disclosure risk. *Res. Off. Statist.*, **1**, 53–67.
Fienberg, S. E. and Makov, U. E. (1998) Confidentiality, uniqueness and disclosure limitation for categorical data. *J. Off. Statist.*, **14**, 385–397.
Fienberg, S. E., Makov, U. E. and Sanil, A. P. (1997) A Bayesian approach to data disclosure: optimal intruder behaviour for continuous data. *J. Off. Statist.*, **13**, 75–89.
Fuller, W. A. (1993) Masking procedures for microdata disclosure limitation. *J. Off. Statist.*, **9**, 383–406.
Goodman, L. A. (1949) On the estimation of the number of classes in a population. *Ann. Math. Statist.*, **20**, 572–579.
Greenberg, B. and Voshell, L. (1990) Relating risk of disclosure for microdata and geographic area size. *Proc. Surv. Res. Meth. Sect. Am. Statist. Ass.*, 450–455.
Greenberg, B. and Zayatz, L. (1992) Strategies for measuring risk in public use microdata files. *Statist. Neerland.*, **46**, 33–48.
Lambert, D. (1993) Measures of disclosure risk and harm. *J. Off. Statist.*, **9**, 313–331.
Marsh, C. (1993) The Samples of Anonymised Records. In *The 1991 Census User's Guide* (eds A. Dale and C. Marsh). London: Her Majesty's Stationery Office.
Marsh, C., Skinner, C., Arber, S., Penhale, B., Openshaw, S., Hobcraft, J., Lievesley, D. and Walford, N. (1991) The case for a sample of anonymized records from the 1991 census. *J. R. Statist. Soc.* A, **154**, 305–340.
Paass, G. (1988) Disclosure risk and disclosure avoidance for microdata. *J. Bus. Econ. Statist.*, **6**, 487–500.
Samuels, S. M. (1998) A Bayesian, species-sampling-inspired approach to the uniques problems in microdata disclosure risk assessment. *J. Off. Statist.*, **14**, 373–383.
Serfling, R. J. (1980) *Approximation Theorems of Mathematical Statistics*. New York: Wiley.
Skinner, C. J. and Holmes, D. J. (1998) Estimating the re-identification risk per record for microdata. *J. Off. Statist.*, **14**, 361–372.
Skinner, C. J., Marsh, C., Openshaw, S. and Wymer, C. (1994) Disclosure control for census microdata. *J. Off. Statist.*, **10**, 31–51.
Willenborg, L. and de Waal, T. (2001) *Elements of Statistical Disclosure Control*. New York: Springer.
Zayatz, L. V. (1991) Estimation of the number of unique population elements using a sample. *Proc. Surv. Res. Meth. Sect. Am. Statist. Ass.*, 369–373.