A Measure of Disclosure Risk for Microdata

C.J. Skinner
University of Southampton

and M.J. Elliot
University of Manchester

Summary

Protection against disclosure is important for statistical agencies releasing microdata files from sample surveys. Estimates of 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 currently used with census data. Second, we show that it may be estimated consistently from sample data without making strong modelling assumptions. This is a surprising finding, in its contrast to the properties of the two 'similar' established measures. As a result, this measure has potentially useful applications to sample surveys. Moreover, we propose a simple variance estimator and show that it is consistent. We also show that the measure and its estimation may be extended to allow for misclassification of identifying variables and to allow for certain complex sampling schemes. We present a numerical study based upon 1991 census data for some 450,000 enumerated individuals in one area of Great Britain. We show that the theoretical results on the properties of the point estimator of the measure of risk and its variance estimator hold to a good approximation for these data.

Key Words: finite population; matching; survey sampling; uniqueness

1. Introduction

Anonymised 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, 1996, 2001). A key element of protection methodology is the assessment of disclosure risk for a file. Assessing disclosure risk usually involves difficult and complex judgements (Lambert, 1993). With the increasing demand for microdata and other trends (Duncan and Pearson, 1991), systematic ways of supporting these judgements by statistical evidence have increasingly been sought. In this paper we consider one common basic form of such evidence: the values of a single measure of disclosure risk, estimated from the data, for alternative possible specifications of the microdata file.

Considerable progress has been made in developing more elaborate forms of evidence to capture more fully the complex nature of potential threats to confidentiality. Duncan and Lambert (1986, 1989), Paass (1988), Lambert (1993), Fuller (1993), Skinner et al.(1994) and Fienberg et al.(1997), among others, have developed statistical modelling frameworks for this purpose. We shall not pursue these more general approaches in this paper, however. We restrict attention to three simple measures of disclosure risk. The first 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 that surprisingly it provides a means of overcoming the inference difficulties for the first two measures. The assessment of disclosure risk using the new measure is the main subject of this paper.

The first measure of disclosure 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 for disclosure risk assessment of census microdata in the U.S.A. and U.K. (Greenberg and Voshell, 1990; Marsh et al., 1991). Bethlehem et al.(1990) provide a systematic discussion, setting set 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 identifying information 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, 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, it may be argued that a more realistic measure of disclosure risk is of the proportion of sample unique records which are population unique, denoted Pr (PU|SU). This is our second measure. It has been used by Statistics Canada to assess disclosure risk in census microdata (Carter et al, 1991), and has been considered further by Skinner et al (1994), Chen and Keller-McNulty (1998), Samuels (1998), Fienberg and Makov (1998) and Elliot et al. (1998)

A potential problem with this second measure is that it is also unrealistic in neglecting the risk presented by records which, although not population unique, are unusual. The third measure of risk again refers to the threat 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 amongst those records in the population 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. The second advantage and the main contribution of this paper is to show in Section 3 that θ may be estimated consistently by a simple point estimator without strong modelling assumptions and that a simple consistent variance estimator is also available. This is a surprising finding since it contrasts with the property discussed in Section 2, that consistent estimation of the 'similar' measures Pr(PU) or Pr(PU|SU) from sample data is problematic in the absence of strong modelling assumptions. For a census, these measures may be calculated by the census office 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 measure from sample data.

It is desirable when measuring disclosure risk to take account of possible risk reducing effects of measurement error in the identifying variables, first because measurement error occurs naturally in surveys and second because it may be used as a masking technique (Fuller, 1993; Fienberg et al, 1997). In Section 4 we show how measurement error may be allowed for in θ. The results in Section 3 assume Bernoulli sampling. The extension of these results to other designs is considered in Section 5. A numerical study of the

properties of the estimation procedures of Section 3 is presented in Section 6. Some concluding remarks are made in Section 7.

2. Measures of Disclosure Risk

In this section we first set out the basic framework and notation. We then define the three measures of disclosure risk, introduced in Section 1. Finally, we compare the three measures, commenting on some advantages of interpretation of the third measure and on difficulties in estimating the first two measures from sample data. The estimation of the third measure will be considered in Section 3.

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 categorical identifying variables, assumed given. The categorical variable formed by cross-classifying all the identifying variables is denoted X, with values denoted 1, ..., J. Each of these values corresponds to a possible combination of values of the identifying variables. For example, if the identifying 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 population unit i. Let the *population frequencies* for the different values of X be denoted

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

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

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

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

$$\sum_{r=1}^{\infty} N_r = J \qquad , \qquad \sum_{r=1}^{\infty} r N_r = N \qquad . \tag{1}$$

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

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

and the *sample frequencies of frequencies* are denoted

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

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 of disclosure risk to be considered is N_1/N, the proportion of units in the population, which are population unique. We write

$$\Pr(PU) = N_1 / N = \sum_j I(F_j = 1) / N$$

as the probability of population uniqueness (PU) for a unit randomly drawn (with equal probabilities) from the population. The second measure of disclosure risk to be considered is given by

$$\Pr(PU \mid SU) = \sum_j I(f_j = 1, F_j = 1) / \sum_j I(f_j = 1).$$

This is the conditional probability that, for a unit randomly drawn from the population, the unit is population unique given that the unit is sample unique.

Finally, the proposed measure of disclosure risk is given by

$$\theta = \sum_j I(f_j = 1) / \sum_j F_j I(f_j = 1) \qquad (2)$$

To interpret θ, suppose that a unit is drawn randomly (with equal probabilities) from the population. Call this the *chosen unit*. Suppose the value of X for the chosen unit is matched 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 $\sum F_j I(f_j = 1)$, the denominator of θ, and the number of these units for which the match is correct is $\sum I(f_j = 1)$, the numerator of θ. Hence we may write

θ =Pr (correct match | unique match)

and interpret θ as the conditional probability that a unique match will be correct.

2.3 **Discussion and Comparison of Three Measures**

Let us first consider the measure Pr(PU). Any population unique microdata record might be viewed as 'risky'. For, if an intruder were able to link such a record to an identifiable unit in the population and know that the unit was population unique then the intruder would know that the link was correct. Thus, one interpretation of Pr(PU) is that it is the expected proportion of sample units which are 'at risk of' disclosure (under sampling with equal inclusion probabilities). One problem with this interpretation as a common measure of risk for all microdata records is that not all microdata records are 'equally

likely' to be population unique. In particular, if a record is not sample unique then it cannot be population unique (assuming no misclassification). Thus the proportion of population uniques among all microdata records (which approximates Pr(PU) under sampling with equal inclusion probabilities) will not exceed Pr (PU|SU), the proportion of population uniques 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 modelling complications, however, which we wish to avoid. We restrict attention to measures which take a single value for the microdata file.

Note that the definition of Pr(PU|SU) (like θ) depends upon the sample s and Pr(PU|SU) is thus not a conventional finite population parameter of the kind considered in survey sampling (Cochran, 1977). The sample-dependent nature of Pr(PU|SU) is, however, natural here since disclosure is conceived of as a property of the sample data rather than the population.

Let us now consider further the adequacy of Pr(PU|SU) as a measure of risk. It seems desirable to interpret Pr(PU|SU) relative to a scenario of attack, according to which an intruder may attempt disclosure. We suggest that the most natural method of attack for which Pr(PU|SU) is relevant is as follows.

Attack Method A: 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 intruder knows that the probability that the selected record is population unique is Pr(PU|SU) and hence that the probability, P, that the selected record belongs to the located unit is at least Pr(PU|SU). For, we may write

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

$$\geq \sum_j I(f_j = 1, F_j = 1) / \sum_j I(f_j = 1) = Pr(PU \mid SU)$$

It may be argued, on the basis of this inequality, that Pr(PU|SU) is an over-optimistic measure of disclosure risk 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. A more appropriate measure under method A is the expression, P, in (3), which may be interpreted as Pr (correct match | unique match) under this method. Note, however, that P is not the same as θ in (2).

Let us turn then to the third measure θ. This will be equal to Pr(correct match | unique match) under the following two methods of attack, which are essentially identical.

Attack Method B: The intruder draws one unit at random from the population and matches this to the microdata. The intruder repeats this process until a unique match is found.

Attack Method B': The intruder takes the whole microdata file and searches through the population at random until a unit is found which uniquely matches one of the microdata records.

Whether θ or P is a more appropriate measure of risk depends on the method of attack. We suggest that method B' is more plausible than method A since the intruder makes fuller use of all the microdata information in the former method B'. 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. Having argued that θ has some advantages of interpretation as a measure of disclosure risk compared to Pr(PU) and Pr(PU|SU), we now comment on the estimation of these two measures from sample data.

The estimation of 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 involved in estimating $J = \sum rN_r$ (Bunge and Fitzpatrick, 1993). A natural approach is to write

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

where E(.) 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 finds the estimator of J can be 'very unreasonable' and the same appears to be the case 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 (4). An alternative 'non-parametric' 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 upon 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 authors proposed an estimator based upon 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) discusses the point estimation of Pr(PU|SU) based on a Poisson-Dirichlet model. Although obtaining some encouraging results, he finds substantial underestimation when the sampling fraction is low and comments (in his Section 6) on the intrinsic difficulties in estimating Pr (PU|SU) from sample data in certain situations. In summary, we suggest that no estimation procedure is currently available which robustly estimates Pr (PU) or 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 estimated without strong modelling assumptions. This is a surprising finding since θ appears to be a 'similar' parameter to the first two measures. A heuristic explanation for this finding is that inference about θ may essentially be achieved by solving only the second of the estimating equations defined by (4) (see Proposition 2) rather than the entire set as required for the estimation of N_1 or J.

3. Estimation of the Proposed Measure

In this section we consider the estimation of θ in (2). We assume that the sample frequencies of frequencies n_r, r=1,2 ... 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 too is the 'parameter' θ of interest (see discussion in section 2.3). 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:

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

To motivate our point estimator of θ, we consider a simulation-based estimator. This is based upon a sample-based analogue of Scenario B in Section 2, referred to here as *Data Intrusion Simulation.*

Data Intrusion Simulation

Repeat the following steps (independently) for k=1, 2, ..., K

Step 1: remove 1 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 removed unit 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 estimator of θ is then the proportion of unique matches which are correct:

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

Step 1 simulates a random drawing of one unit from the population, as in Scenario B, since the sample units are assumed to be drawn with equal probabilities. Likewise, Step 2 simulates the fact that the population unit selected by the intruder will be included in the sample with probability π. The estimator $\hat{\theta}(K)$ is formed from the usual 'analogue' principle that a sample quantity is a natural estimator of the corresponding population quantity. This principle does not generate a sensible estimator of Pr(PU) or Pr(PU|SU), however, and so it is natural to be sceptical initially as to whether $\hat{\theta}(K)$ will be a sensible estimator 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: $\hat{\theta}(K) \to \hat{\theta}$ a.s. (with respect to the randomisation in the simulation), where

$$\hat{\theta} = \pi n_1 / \left[\pi n_1 + 2(1-\pi) n_2 \right] \qquad (7)$$

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

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 estimator has been introduced only to motivate $\hat{\theta}$ and possible extensions (see Section 4).

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

$$\hat{\theta} = n_1 / \left[n_1 + 2(1-\pi) n_2 / \pi \right] \quad , \quad \theta = n_1 / \left[n_1 + \sum_j (F_j - 1) I(f_j = 1) \right].$$

Hence $\hat{\theta}$ will be a sensible estimator of θ if n_2 is a sensible estimator of $\pi \sum (F_j - 1) I(f_j = 1) / \left[2(1-\pi) \right]$. The unbiasedness of the latter estimator with respect to the Bernoulli sampling is implied by the following result.

Proposition 2: $$E\left[2(1-\pi) n_2 / \pi \right] = E\left[\sum_j (F_j - 1) I(f_j = 1) \right] \qquad (8)$$

In order to demonstrate the consistency of $\hat{\theta}$ as an estimator of θ, we consider an asymptotic framework defined by a sequence of populations indexed by increasing J.

In practice J will usually be large, for example it is 4.3×10^6 in the application in Section 6 (including catgegories with zero counts). We conceive of an increase in J as corresponding to an increase in the number of identifying variables, associated with an increase in the population size $(N = \sum^J F_j)$ in such a way that the maximum value of F_j remains bounded. The expected sample size E(n) = πN (from (5)) will increase in proportion to the population size (with π treated as fixed).

Proposition 3: Under Bernoulli sampling with sampling fraction π, $0<\pi<1$, and assuming that the F_j are bounded above, we have $\hat{\theta}-\theta = o_p(1)$, $\left(\hat{\theta}-\theta\right)/v^{½} \rightarrow N(0,1)$ as $J\rightarrow\infty$

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

$$c = \left[\sum F_j\pi(1-\pi)^{F_j-1}\right]/\left[\sum F_j^2\pi(1-\pi)^{F_j-1}\right]^2$$

Thus, under the given asymptotic framework, $\hat{\theta}$ is consistent for θ, in the sense that $\hat{\theta}-\theta$ converges in probability to zero, and $\hat{\theta}-\theta$ is asymptotically normal. A simple consistent estimator of the variance of $\hat{\theta}-\theta$ is given by

$$\hat{v} = \hat{\theta}^2\,\frac{2(1-\pi)\left[3(1-\pi)n_3+(2-\pi)n_2\right]}{\left[\pi n_1+2(1-\pi)n_2\right]^2} \qquad (9)$$

Its consistency for v is now demonstrated.

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

A corollary of Propositions 3 and 4 is, from Slutsky's Theorem, that

$\left(\hat{\theta}-\theta\right)/\hat{v}^{½} \rightarrow N(0,1)$ as $J\rightarrow\infty$.

For the disclosure control application, an agency might adopt a conservative approach by considering the upper bound of a one-sided confidence interval $\hat{\theta} + z_{1-\alpha}\hat{v}^{½}$, where $z_{1-\alpha}$ is the $100z_{1-\alpha}$ percentage point of the standard normal distribution, and requiring this bound to be sufficiently low, say below 0.1.

4. **Misclassification**

In this section we extend the definition of $\theta = \Pr$ (correct match | unique match) to accommodate misclassification and show that the Data Intrusion Simulation approach of Section 3 may be naturally extended to estimate θ consistently. To allow for misclassification, we now let X denote the combination of values of the identifying variables as recorded in the microdata and $\tilde{X}$ denote the corresponding variable as measured by a potential intruder using external information. We say that *misclassification* occurs for unit i if $X_i \neq \tilde{X}_i$. We do not assume that either X or $\tilde{X}$ measures 'truth'. They simply reflect two ways of classifying the same quantity. In particular, it is possible that either X or $\tilde{X}$ is subject to measurement error and that X is subject to deliberate perturbation as a means of disclosure control.

By analogy with the definitions of F_j, f_j, N_r and n_r in Section 2.1, we let

$$\tilde{F}_j = \sum_{i \in U} I\left(\tilde{X}_i = j\right),\ \tilde{f}_j = \sum_{i \in s} I\left(\tilde{X}_i = j\right),\quad \tilde{N}_r = \sum_j I\left(\tilde{F}_j = r\right), \tilde{n}_r = \sum_j I\left(\tilde{f}_j = r\right).$$

We define θ again as Pr (correct match | unique match) under attack method B of Section 2.3, that is

$$\theta = \Pr \text{ (correct match | unique match)} = \frac{\sum_{i \in s} I\left(f_{X_i} = 1, \tilde{X}_i = X_i\right)}{\sum_j \tilde{F}_j I\left(f_j = 1\right)} \qquad (10)$$

Note that θ is the same as in (2) if there is no misclassification. In order to estimate θ we assume that misclassification takes place according to a random mechanism in which

$$\Pr\left(\tilde{X}_i = j^* \mid X_i = j\right) = M_{jj^*} \qquad j = 1,...J \quad , \quad j^* = 1,...J \; , \; i \in U \tag{11}$$

where the matrix $M = \left[M_{jj^*} \right]$ is a $J \times J$ misclassification matrix. We further assume that M is known. In practice, an agency will not know M exactly, but may conduct a sensitivity analysis for various plausible values of M (c.f. Kuha and Skinner 1997). In order to obtain a point estimator of $\hat{\theta}$ of θ we add a step to the Data Intrusion Simulation of Section 3.

Data Intrusion Simulation under Misclassification

Repeat the following steps (independently) for k=1, ..., K.

Step 1: remove 1 unit at random from the sample;

Step 2: determine the value of $\tilde{X}$ for this unit randomly using M, that is set $\tilde{X} = j^*$ with probability M_{jj^*}, where j is the unit's value of X;

Step 3: copy the unit back into the sample with probability π (keeping its original value of X);

Step 4: record whether the value $\tilde{X}$ of the removed unit matches uniquely the value of X for a sample unit ($R_{uk} = 1$ if so) and, if so, whether this match is correct ($R_{ck} = 1$ if so).

The resulting estimator of θ is again given by $\hat{\theta}(K)$ in (6). As before we may obtain a closed form expression, $\hat{\theta}$, as the limit of $\hat{\theta}(K)$ as $K \to \infty$. Events 1 and 2 in the proof of Proposition 1 now require correct classification at Step 2 (which occurs with probability M_{jj}) for the match to be unique. A third possible event when a unique match arises must also be considered, that in Step 2 X is misclassified to a value j which

corresponds to a sample unique in the microdata at Step 3. This event occurs with probability

$$A = \sum_{i \in s} \sum_{j=1}^{J} M_{X_i j} I(X_i \neq j) I(f_j = 1) / n$$

Following an analogous argument to the proof of Proposition 1, we obtain

$$\hat{\theta} = \frac{\pi \sum_j I(f_j = 1) M_{jj}}{\pi \sum_j I(f_j = 1) M_{jj} + 2(1-\pi) \sum_j I(f_j = 2) M_{jj} + nA} \quad (12)$$

Note that, in the absence of misclassification, $M_{jj} = 1$, $A = 0$ and $\hat{\theta}$ reduces to the expression in (7). The expression for A reduces in general to

$$A = \sum_j \left[E_M(\tilde{f}_j) - M_{jj} \right] I(f_j = 1) / n, \quad (13)$$

where $E_M(.)$ denotes the expected value with respect to the misclassification mechanism. The consistency of $\hat{\theta}$ for θ is now shown.

Proposition 5: $\hat{\theta} - \theta = o_p(1)$, under the probability distribution induced by both the Bernoulli sampling and the misclassification mechanism in (11), where $\hat{\theta}$ and θ are defined in (12) and (10) respectively and the assumptions of Proposition 3 hold.

5. 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.

5.1. Simple Random Sampling Without Replacement

Under simple random sampling of size n, the binomial distribution of f_j in (5) is replaced by the hypergeometric distribution with parameters (N, n, F_j). We define $\hat{\theta}$ as in (7)

with $\pi = n/N$. Proposition 2 no longer holds exactly but it is straightforward to verify that it does so approximately in an asymptotic framework where $n \to \infty$, $N \to \infty$, $n/N \to \pi$ (fixed) and the F_j are bounded. Using this result, we may extend the argument in Proposition 3 to show that $\hat{\theta}$ remains consistent for θ under simple random sampling within this framework. This follows more directly by noting that, under this asymptotic framework, the hypergeometric distribution for f_j is approximated by the binomial distribution in (5) and the f_j are approximately independent $(j = 1,...,J)$.

5.2 **Proportionate Stratification**

Suppose now that the population consists of H strata of sizes $N_1,...,N_H$ and that independent simple random samples of sizes $n_1,...,n_H$ 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 $\sum_h f_{hj} = f_j, \sum_h F_{hj} = F_j$), the distribution of f_{hj} is now hypergeometric with parameters (N_h, n_h, F_{hj}). We assume $n_h/N_h = \pi$ and define $\hat{\theta}$ as in (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 $(h = 1,...,H,\ j = 1...J)$ then, as in Section 5.1, the f_{hj} will be approximately independently binomially distributed with parameters (F_{hj}, π). Under independent sampling in different strata, the distribution of f_j will again be as in (5) and the consistency of $\hat{\theta}$ holds.

5.3 **Unequal Probability Sampling**

Suppose now that the sample inclusion probabilities π_i of different population units may now 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 (7), since it depends on π, an assumed common inclusion probability. It does not appear to be straightforward to modify $\hat{\theta}$ to achieve consistent estimation of θ under general unequal probability designs. Perhaps the most natural modification of the Data Intrusion Simulation in Section 3 would be to modify Step 1 to remove the sample unit i with probability $\pi_i^{-1} / \sum_s \pi_i^{-1}$ (in order to mimic selecting one unit with equal probability from the population) and to modify Step 2 to copy this unit back into the sample with probability π_i. The properties of the resulting estimator, $\hat{\theta}$, require further research.

5.4 **Cluster Sampling**

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

Example Suppose the population is partitioned into clusters of size 1 or 2 and cluster sampling is employed with equal inclusion probability π. Suppose 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 2 and f_j=0 or F_j for all j. It follows from (2) that θ =1. But n_2 will not in general be zero 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 estimator of θ in this example.

To obtain a heuristic guide to the impact of cluster sampling, note that for $\hat{\theta}$ to be a reasonable estimator of θ, the ratios $\Pr(f_j = 2) / \Pr(f_j = 1)$ should roughly follow that for the binomial distribution of (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 the case for many social surveys. The estimator $\hat{\theta}$ will be most harmed when the cluster sampling has a clear direct effect on the $\Pr(f_j = 1)$ 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 $\Pr(f_j = 2)/\Pr(f_j = 1)$ relative to what would be expected under Bernoulli sampling of individuals.

6. Numerical Study

The aim of this section is to provide some numerical evidence on the properties of the procedures described in Section 2 and 3. 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), country of birth (42). This choice of identifying variables was based upon 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 similar results to those presented here.

Samples were drawn from this population using systematic sampling of 1 in L units for L=10, 20 and 50 i.e. with π=0.1,0.05 and 0.02, within geographical strata.

The stratification follows that used to draw the individual Sample of Anonymised 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 $\hat{v}^{½}$ by enumerating all L possible samples.

Table 1 contains the means and standard deviations across the L systematic samples for $\theta, \hat{\theta}$ and $\hat{v}^{½}$ defined in (2), (7) and (9) respectively. Considering the measure of disclosure risk 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 estimator $\hat{\theta}$, we may 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 $\hat{\theta}$ (s.d. $\left(\hat{\theta}-\theta\right)$) 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 π=0.02, 0.05 and 0.10 respectively so $\hat{\theta}$ is a fairly stable estimator of θ here. The estimator $\hat{v}^{½}$ of the standard error of $\hat{\theta}$ does appear to be approximately unbiased. There is a slight upward bias (implying $\hat{v}$ is a conservative variance estimator) arising perhaps from the stratified systematic design reducing the variance of $\hat{\theta}-\theta$. The

coefficient variation of $\hat{v}^{½}$ is 7.6%, 8.7% and 3.2% for π = 0.02, 0.05 and 0.10 respectively and so $\hat{v}$ is also a fairly stable estimator of the variance of $\hat{\theta} - \theta$ here.

7. Concluding Remarks

In this paper, we have proposed a measure of disclosure risk for sample microdata based on the probability that an observed match between a microdata record and an external record is correct. We have argued that this measure has a clear and useful interpretation and that it may be estimated simply from sample microdata. The proposed approach to estimation has been shown to possess desirable theoretical properties and to perform well in a numerical study based upon census data for a population of 450,000 individuals from an area in Great Britain.

The proposed measure might be used by a statistical agency trying to choose 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 area of residence or 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 $\hat{v}^{½}$ 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 absolute way, for example by requiring that $\hat{\theta}$ (or $\hat{\theta}$ + 2.3 $\hat{v}^{½}$) may not exceed some specified probability, for example 0.1.

We have shown that our approach may also be extended to allow for misclassification of potential identifying variables. Empirical investigation of this extension remains to be undertaken. 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 multi-stage 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 multi-stage design, where the multi-stage units are not strongly related to the categories of X defined by potential identifying variables.

Appendix : Proofs

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

$$\Pr(R_{uk}=1)=\left[\pi n_1+2(1-\pi)n_2\right]/n \quad , \quad \Pr(R_{ck}=1\,|\,R_{uk}=1)=\hat{\theta} \qquad \text{(A.1)}$$

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

Proof of Proposition 2: It follows from (5) that both sides of (8) equal

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

Outline Proof of Proposition 3: We may write $\hat{\theta}-\theta=g(T_J)$, where $T_J=\sum Y_j$ and $Y_j=\left[I(f_j=1),I(f_j=2),F_jI(f_j=1)\right]'$. The assumptions of the proposition are sufficient for a central limit theorem for the independent random vectors Y_1, ..., Y_J, giving $V_J^{-½}\left[T_J-\mu_J\right]\to N(0,1)$ as $J\to\infty$ where $\mu_J=E(T_J)$ and $V_J=\mathrm{var}(T_J)$ are defined with respect to the binomial distributions in (5). It follows by the delta method that

$$\left[\hat{\theta}-\theta-g(\mu_J)\right]/v^{½}\to N(0,1)$$

where $v = \mathrm{var}\left[\nabla'\left(T_J - \mu_J\right)\right]$ 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

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

$$= \mu_{J1}\mu_{J3}^{-2}\left[-1, -\phi, 1\right] \text{ , using again the fact that } \mu_{J3} = \mu_{J1} + \phi\mu_{J2}.$$

Hence $v = \left(\mu_{J1}\mu_{J3}^{-2}\right)^2 \mathrm{var}\left(-T_{J1} - \phi T_{J2} + T_{J3}\right)$, where $T_J = (T_{J1}, T_{J2}, T_{J3})'$. (A.2)

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

Using Proposition 2, we have

$$\mathrm{var}\left(-T_{J1} - \phi T_{J2} + T_{J3}\right) = \sum_j E\left[\left(F_j - 1\right) I\left(f_j = 1\right) - \phi I\left(f_j = 2\right)\right]^2$$

$$= \sum_j E\left[\left(F_j - 1\right)^2 I\left(f_j = 1\right)\right] + \phi^2 \Pr\left(f_j = 2\right)$$

$$= \sum_j \left(F_j - 1\right)^2 F_j \pi (1-\pi)^{F_j - 1} + \phi^2 F_j \left(F_j - 1\right)\pi^2 (1-\pi)^{F_j - 2}/2$$

$$= \sum_j F_j \left(F_j - 1\right)(1-\pi)^{F_j - 1}\left[\left(F_j - 1\right)\pi + 2(1-\pi)\right]$$

$$= \sum_J F_j \left(F_j - 1\right)\left(\pi F_j - 3\pi + 2\right)(1-\pi)^{F_j - 1} \quad \text{(A.3)}$$

Finally the expression for v in the Proposition is obtained from (A.2) and (A.3) by noting

$$\mu_{J1} = \sum_j F_j \pi (1-\pi)^{F_j - 1} \quad , \quad \mu_{J3} = \sum_j F_j^2 \pi (1-\pi)^{F_j - 1}.$$

Outline Proof of Proposition 4: Note that $\hat{\theta} = \mu_{J1}/\mu_{J3} + o_p(1)$,

$\left[\pi n_1 + 2(1-\pi)n_2\right]/J = \pi\mu_{J3}/J + o_P(1)$ and

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

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 $c = \mu_{J1}/\mu_{J3}^2$.

Outline Proof of Proposition 5: Note first that by taking expectations of the numerator and denominators of (10) with respect to the misclassification mechanism and using independence between j we have

$$\theta = \sum_j I(f_j = 1)M_{jj} / \sum_j I(f_j = 1)E_M(\tilde{F}_j) + o_p(1) \qquad \text{(A.4)}$$

By comparing expressions (12) and (A.4) it is sufficient to show that

$$\sum_j I(f_j = 1)M_{jj}/n + \left[2(1-\pi)/\pi\right]\sum_j I(f_j = 2)M_{jj}/n + A/\pi$$

$$= \sum_j I(f_j = 1)E_M(\tilde{F}_j)/n + o_p(1) \qquad \text{(A.5)}$$

This may be shown using the following results

$$E\left[I(f_j = 1)M_{jj}\right] = F_j\pi(1-\pi)^{F_j-1}M_{jj}$$

$$E\left[I(f_j = 2)M_{jj}\right] = 0.5F_j(F_j-1)\pi^2(1-\pi)^{F_j-2}M_{jj}$$

$$E[A] = E\left[\sum_j\left(\sum_{j^*\neq j} f_{j^*}M_{j^*j}\right)I(f_j = 1)\right]/n = \sum_j \pi\left[E_M(\tilde{F}_j) - F_jM_{jj}\right]F_j\pi(1-\pi)^{F_j-1}/n$$

$$E\left[I(f_j = 1)E_M(\tilde{F}_j)\right] = F_j\pi(1-\pi)^{F_j-1}E_M(\tilde{F}_j)$$

where the expectations E(.) are with respect to the Bernoulli sampling.

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. *Statistica Neerlandica,* **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. Social Survey Methods Division Report, Statistics Canada.

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

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

Duncan, G.T. and Lambert, D. (1986) Disclosure limited data dissemination. *J.Am.Statist. Ass.*, **81**, 10-28.

Duncan, G.T. and Lambert, D. (1989) The risk of disclosure for microdata. *J.Bus.Econ. Statist.*, **7**, 207-217.

Duncan G.T. and Pearson, R.W. (1991) Enhancing access to the microdata while protecting confidentiality: prospects for the future (with discussion). *Statist.Sci.*, **6**, 219-239.

Elliot, M.J. (2000) DIS: a new approach to the measurement of statistical disclosure risk. *Int.J.Risk Management*, **2**(4), 39-48.

Elliot, M.J. and Dale, A. (1999) Scenarios of attack : the data intruder's perspective on statistical disclosure risk. *Netherlands Official 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. *Research in Official Statist.*, **1**, 53-67.

Fienberg, S.E. and Makov, U.E. (1998) Confidentiality, uniqueness and disclosure limitation for categorical data. *J. Official 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.Official Statist.*, **13**, 75-89.

Fuller, W.A. (1993) Masking procedures for microdata disclosure limitation. *J.Official 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.Sect.Survey Res.Meth.*, Am.Statist.Ass., 450-455.

Greenberg, B. and Zayatz, L. (1992) Strategies for measuring risk in public use microdata files. *Statistica Neerlandica*, **46**, 33-48.

Kuha, J.T. and Skinner, C.J. (1997) Categorical data analysis and misclassification. In L. Lyberg, P. Biemer, M. Collins, E. De Leeuw, C. Dippo, N. Schwarz, and D. Trewin eds. *Survey Measurement and Process Quality*, New York: Wiley, 633-670.

Lambert, D. (1993) Measures of disclosure risk and harm. *J.Official Statist.*, **9**, 313-331.

Marsh, C. (1993) The Samples of Anonymised Records. In A. Dale and C. Marsh eds. *The 1991 Census User's Guide,* 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 anonymised records from the 1991 Census. *J.R.Staistt.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.Official Statist.*, **14**, 373-383.

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

Skinner, C.J., Marsh, C., Openshaw, S., and Wymer, C. (1994) Disclosure control for census microdata. *J.Official Statist.*, **10**, 31-51.

Willenborg, L., and de Waal, T. (1996) *Statistical Disclosure Control in Practice.* New York: Springer-Verlag.

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

Table 1. Means and Standard Deviations of Different Quantities Across All Possible Systematic Samples of a Specified Sampling Fraction from Census Population of 450,000 individuals

		Sampling Fraction π		
		0.02	0.05	0.10
Risk Measure, θ	Mean	0.0426	0.1047	0.1985
	s.d.	0.0012	0.0051	0.0027
Estimator, $\hat{\theta}$	Mean	0.0429	0.1055	0.1990
	s.d.	0.0020	0.0058	0.0045
Error, $\hat{\theta} - \theta$	Mean	0.0004	0.0008	0.0005
	s.d.	0.0025	0.0051	0.0061
S.E. Estimator, $\hat{v}^{1/2}$	Mean	0.0028	0.0052	0.0072
	s.d.	0.0002	0.0004	0.0002