Sample Surveys: Design, Methods and Applications, Vol. 29A
ISSN: 0169-7161

DOI: 10.1016/S0169-7161(08)00015-1

Statistical Disclosure Control for Survey Data

Chris Skinner

1. Introduction

1.1. The problem of statistical disclosure control

Survey respondents are usually provided with an assurance that their responses will be treated confidentially. These assurances may relate to the way their responses will be handled within the agency conducting the survey or they may relate to the nature of the statistical outputs of the survey as, for example, in the "confidentiality guarantee" in the United Kingdom (U.K.) National Statistics Code of Practice (National Statistics, 2004, p. 7) that "no statistics will be produced that are likely to identify an individual." This chapter is concerned with methods for ensuring that the latter kinds of assurances are met. Thus, in the context of this chapter, *statistical disclosure control* (SDC) refers to the methodology used, in the design of the statistical outputs from the survey, for protecting the confidentiality of respondents' answers. Methods relating to the first kind of assurance, for example, computer security and staff protocols for the management of data within the survey agency, fall outside the scope of this chapter.

There are various kinds of *statistical outputs* from surveys. The most traditional are tables of descriptive estimates, such as totals, means, and proportions. The release of such estimates from surveys of households and individuals have typically not been considered to represent a major threat to confidentiality, in particular because of the protection provided by sampling. Tabular outputs from the kinds of establishment surveys conducted by government have, however, long been deemed risky, especially because of the threat of disclosure of information about large businesses in cells of tables which are sampled with a 100% sampling fraction. SDC methods for such tables have a long history and will be outlined in Section 2.

Although the traditional model of delivering all the estimates from a survey in a single report continues to meet certain needs, there has been increasing demand for more flexible survey outputs, often for multiple users, where the set of population parameters of interest is not prespecified. There are several reasons why it may not be possible to prespecify all the parameters. Data analysis is an iterative process, and what analyses are of most interest may only become clear after initial exploratory analyses of the data. Moreover, given the considerable expense of running surveys, it is natural for many

commissioners of surveys to seek to facilitate the use of the data by multiple users. But it is usually impossible to prespecify all possible users and their needs in advance. A natural way to provide flexible outputs from a survey to address such needs is to make the survey microdata available, so that users can carry out the statistical analyses that interest them.

However, the release of such microdata raises serious confidentiality protection issues. Of course, statistical analyses of survey data do not require that the identities of the survey units are known. Names, addresses, and contact information for individuals or establishment can be stripped from the data to form an *anonymized* microdata file. The problem, however, is that such basic anonymization is often insufficient to protect confidentiality, and therefore, it is necessary to use one of a range of alternative approaches to SDC and this will be discussed further in Section 3.

1.2. Concepts of confidentiality, disclosure, and disclosure risk

To be precise about what is meant by "protecting confidentiality" requires discussion of definitions. These usually involve the notion of a hypothetical *intruder* who might seek to breach confidentiality. There are thus three key parties: (1) the *respondent* who provides the data, (2) the *agency* which collects the data, releases statistical outputs, and designs the SDC strategy, and (3) the hypothetical *intruder* who has access to these outputs and seeks to use them to disclose information about the respondent. One important notion of disclosure is *identity disclosure* or *identification,* which would occur if the intruder linked a known individual (or other unit) to an individual microdata record or other element of the statistical output. Another important notion is *attribute disclosure*, which would occur if the intruder could determine the value of some survey variable for an identified individual (or other unit) using the statistical output. More generally, *prediction disclosure* would occur if the intruder could predict the value of some survey variable for an identified individual with some uncertainty. When assessing the potential for disclosure for a particular statistical output, it is usual to refer to the *disclosure risk*. This might be defined as the probability of disclosure with respect to specified sources of uncertainty. Or the term might be used loosely to emphasize not only the uncertainty about potential disclosure but also the potential harm that might arise from disclosure (Lambert, 1993). The *confidentiality* of the answers provided by a respondent might be said to be protected if the disclosure risk for this respondent and the respondent's answers is sufficiently low. In this chapter, disclosure risk is discussed in more detail in Sections 2 and 3. For further discussion of definitions of disclosure, see Duncan and Lambert (1986, 1989) and Skinner (1992).

1.3. Approaches to protecting confidentiality

If the disclosure risk is not deemed to be sufficiently low, then it will be necessary to use some method to reduce the risk. There are broadly two approaches, which are referred to here as *safe setting* and *safe data* (Marsh et al., 1994). The safe setting approach imposes restrictions on the set of possible users of the statistical output and/or on the ways that the output can be used. For example, users might be required to sign a licensing agreement or might only be able to access microdata by visiting a secure laboratory or by submitting

requests remotely (National Research Council, 2005). The safe data approach, on the other hand, involves some modification to the statistical output. For example, the degree of geographical detail in a microdata file from a national social survey might be limited so that no area containing less than 100,000 households is identified. In this chapter, we focus on the safe data approach and generally refer to methods for modifying the statistical output as SDC methods.

1.4. SDC methods, utility, and data quality

SDC methods vary according to the form of the statistical output. Some simple approaches are as follows:

- *Reduction of detail*, for example, the number of categories of a categorical variable might be reduced in a cross-classified table or in microdata.
- *Suppression*, for example, the entry in a table might be replaced by an asterisk, indicating that the entry has been suppressed for confidentiality reasons.

In each of these cases, the SDC method will lead to some *loss of information* for the user of the statistical output. Thus, the method will reduce the number of population parameters for which a user can obtain survey estimates. Other kinds of SDC methods might not affect the number of parameters which can be estimated but may affect the *quality* of the estimates that can be produced. For example, if *random noise* is added to an income variable to protect confidentiality, then this may induce bias or variance inflation in associated survey estimates. The general term *utility* may be used to cover both the information provided by the statistical outputs, for example, the range of estimates or analyses which can be produced, and the quality of this information, for example, the extent of errors in these estimates. It should, of course, be recognized that survey data are subject to many sources of error, even prior to the application of SDC methods, and the impact of SDC methods on data quality therefore needs to be considered in this context.

Generally, utility needs to be considered from the perspective of a *user* of the statistical outputs, who represents a key fourth party to add to the three parties referred to earlier: the respondent, the agency, and the intruder.

1.5. SDC as an optimization problem: the risk-utility trade-off

The key challenge in SDC is how to deal with the trade-off between disclosure risk and utility. In general, the more the disclosure risk is reduced by an SDC method, the lower will be the expected utility of the output. This trade-off may be formulated as an optimization problem. Let D be the (anonymized) survey data and let $f(D)$ be the statistical output, resulting from the use of an SDC method. Let $R[f(D)]$ be a measure of the disclosure risk of the output, and let $U[f(D)]$ be a measure of the utility of the output. Then, the basic challenge of SDC might be represented as the constrained optimization problem:

$$\text{for given } D \text{ and } \varepsilon, \text{ find an SDC method, } f(.), \text{ which maximizes } U[f(D)], \text{ subject to } R[f(D)] < \varepsilon.$$

The elements of this problem need some clarification:

$f(.)$: the *SDC method*—a wide variety of these have been proposed and we shall refer to some of these in this chapter;

$R(.)$: the *disclosure risk function*—we shall discuss ways in which this function may be defined; this is certainly not straightforward, for example, because of its dependence on assumptions about the intruder and because of the challenge of combining the threats of disclosure for multiple respondents into a scalar function;

$U(.)$: the *utility function*—this will also not be straightforward to specify as a scalar function, given the potential multiple uses of the output;

ε : the *maximum acceptable risk*—in principle, one might expect the agency to provide this value in the light of its assurances to respondents. However, in practice, agencies find it very difficult to specify a value of ε, other than zero, that is, no disclosure risk. Unfortunately, for most definitions of disclosure risk, the only way to achieve no disclosure risk is by not releasing any output and this is rarely a solution of interest!

Given these difficulties in specifying $R(.)$ and $U(.)$ as scalar functions and in specifying a value for ε, the above optimization problem serves mainly as conceptual motivation. In practice, different SDC methods can be evaluated and compared by considering the values of alternative measures of risk and utility. For given measures of each, it can sometimes be useful to construct an RU map (Duncan et al., 2001), where a measure of risk is plotted against a measure of utility for a set of candidate SDC methods. The points on this map are expected to display a general positive relationship between risk and utility, but one might still find that, for given values of risk, some methods have greater utility than others and thus are to be preferred. This approach avoids having to assume a single value of ε.

2. Tabular outputs

2.1. Disclosure risk in social surveys and the protection provided by sampling

The main developments in SDC methods for tabular outputs have been motivated by the potential risks of disclosure arising when 100% sampling has been used, such as in censuses or in administrative data. Frequency tables based upon such data sources may often include small counts, as low as zero or one, for example, in tables of numbers of deaths by area by cause of death. Such tables might lead to identity disclosure, for example, if it is public knowledge that someone has died, then it might be possible to identify that person as a count of one in a table of deaths using some known characteristics of that person. Attribute disclosure might also occur. For example, it might be possible to find out the cause of the person's death if the table cross-classifies this cause by other variables potentially known to an intruder.

In social surveys, however, the use of sampling greatly reduces the risks of such kinds of disclosure for two reasons. First, the presence of sampling requires different kinds of statistical outputs. Thus, the entries in tables for categorical variables tend to

be weighted proportions (possibly within domains defined by rows or columns) and not unweighted sample counts. Even if a user of the table could work out the cell counts (e.g., because the survey uses equal weights and the sample base has been provided), the survey agency will often ensure that the published cells do not contain very small counts, where the estimates would be deemed too unreliable due to sampling error. For example, the agency might suppress cell entries where the sample count in the cell falls below some threshold, for example, 50 persons in a national social survey. This should prevent the kinds of situations of most concern with 100% data. Sometimes, agencies use techniques of small area estimation (see Chapters 31 and 32) in domains with small sample counts and these techniques may also act to reduce disclosure risk.

Second, the presence of sampling should reduce the precision with which an intruder could achieve predictive disclosure. For example, suppose that an intruder could find out from a survey table that, among 100 respondents falling into a certain domain, 99 of them have a certain attribute and suppose that the intruder knows someone in the population who falls into this domain. Then, the intruder cannot predict that this person has the attribute with probability 0.99, since this person need not be a respondent and prediction is subject to sampling uncertainty. This conclusion depends, however, on the identities of the survey respondents being kept confidential by the agency, preventing the intruder knowing whether the known person is a respondent, referred to as *response knowledge* by Bethlehem et al. (1990). In general, it seems very important that agencies do adopt this practice since it greatly reduces disclosure risk while not affecting the statistical utility of the outputs. In some exceptional cases, it may be difficult to achieve this completely. For example, in a survey of children it will usually be necessary to obtain the consent of a child's parent (or other adult) in order for the child to take part in the survey. The child might be assured that their responses will be kept confidential from their parent. However, when examining the outputs of the survey, the parent (as intruder) would know that their child was a respondent.

For the reasons given above, disclosure will not generally be of concern in the release of tables of estimates from social surveys, where the sample inclusion probabilities are small (say never exceeding 0.1). See also Federal Committee on Statistical Methodology (2005, pp. 12–14).

2.2. *Disclosure risk in establishment surveys*

A common form of output from an establishment survey consists of a table of estimated totals, cross-classified by characteristics of the establishment. Each estimate takes the form $\hat{Y}_c = \sum_s w_i I_{ci} y_i$, where w_i is the survey weight, I_{ci} is a 0–1 indicator for cell c in the cross-classification, and y_i is the survey variable for the ith establishment in the sample s. For example, y_i might be a measure of output and the cells might be formed by cross-classifying industrial activity and a measure of size.

The relevant definition of disclosure in such a setting will often be a form of prediction disclosure. Prediction disclosure for a specific cell c might be defined under the following set-up and assumptions:

– the intruder is one of the establishments in the cell which has the aim of predicting the value y_i for one of the other establishments in the cell or, more generally, the

intruder consists of a *coalition* of m of the N_c establishments in the cell with the same predictive aim;
- the intruder knows the identities of all establishments within the cell (since, e.g., they might represent businesses competing in a similar market).

Given such assumptions, prediction disclosure might be said to occur if the intruder is able to predict the value y_i with a specified degree of precision. To clarify the notion of precision, we focus in the next subsection on the important case where the units in the cell all fall within completely enumerated strata. Thus, $w_i = 1$ when $I_{ci} = 1$ so that $\hat{Y}_c = \sum_{U_c} y_i$, where U_c is the set of all establishments in cell c and N_c is the size of U_c. In this case, the intruder faces no uncertainty due to sampling and this might, therefore, be treated as the worst case.

2.2.1. Prediction disclosure in the absence of sampling

In the absence of sampling, prediction is normally considered from a deterministic perspective and is represented by an interval (between an upper and lower bound) within which the intruder knows that a value y_i must lie. The precision of prediction is represented by the difference between the true value and one of the bounds. It is supposed that the intruder undertakes prediction by combining prior information with the reported value $\hat{Y}_c$.

One approach to specifying the prior information is used in the *prior-posterior rule* (Willenborg and de Waal, 2001), also called the *pq* rule, which depends upon two constants, p and q, set by the agency. The constant q is used to specify the precision of prediction based upon the prior information alone. Under the *pq* rule, it is assumed that intruder can infer the y_i value for each establishment in the cell to within $q\%$. Thus, the agency assumes that, prior to the table being published, the intruder could know that a value y_i falls within the interval $[(1 - q/100)y_i, (1 + q/100)y_i]$. The combination of this prior information with the output $\hat{Y}_c = \sum_{U_c} y_i$ can then be used by the intruder to obtain sharper bounds on a true value. For example, let $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(N_c)}$ be the order statistics and suppose that the intruder is the establishment with the second largest value, $y_{(N_c-1)}$. Then, this intruder can determine an upper bound for the largest value $y_{(N_c)}$ by subtracting its own value $y_{(N_c-1)}$ together with the sum of the lower bounds for $y_{(1)}, \ldots, y_{(N_c-2)}$ from $\hat{Y}_c$. The precision of prediction using this upper bound is given by the difference between this upper bound and the true value $y_{(N_c)}$, which is $(q/100)\sum_{i=1}^{N_c-2} y_{(i)}$. This cell would be called *sensitive* under the *pq* rule, that is, judged *disclosive*, if this difference was less than $p\%$ of the true value, that is, if

$$(p/100)y_{(N_c)} - (q/100)\sum_{i=1}^{N_c-2} y_{(i)} > 0. \tag{1}$$

The expression on the left-hand side of (1) is a special case of a *linear sensitivity measure*, which more generally takes the form $R_c = \sum_{i=1}^{N_c} a_i y_{(i)}$, where the a_i are specified weights. The cell is said to be sensitive if $R_c > 0$. In this case, prediction disclosure would be deemed to occur. A widely used special case of the *pq* rule is the *p% rule*, which arises from setting $q = 100$, that is, no prior information is assumed. Another commonly used linear sensitivity measure arises with the (n, k) or *dominance rule*. See Willenborg and de Waal (2001), Cox (2001), Giessing (2001), and Federal Committee on Statistical Methodology (2005) for further discussion.

2.2.2. Prediction disclosure in the presence of sampling

More generally, all cell units may not be completely enumerated. In this case, $\hat{Y}_c$ will be subject to sampling error and, in general, this will lead to additional disclosure protection, provided that the intruder does not know whether other establishments (other than those in the coalition) are sampled or not. The definition of risk in this setting appears to need further research. Willenborg and de Waal (2001, Section 6.2.5) presented some ideas. An alternative model-based stochastic approach might assume that before the release of the table, the prior information about the y_i can be represented by a linear regression model depending upon publicly available covariate values x_i with a specified residual variance. The predictive distribution of y_i given x_i could then be updated using the known value(s) of y_i for the intruder and the reported $\hat{Y}_c$, which might be assumed to follow the distribution $\hat{Y}_c \sim N[Y_c, v(\hat{Y}_c)]$, where $v(\hat{Y}_c)$ is the reported variance estimate of $\hat{Y}_c$. Prediction disclosure could then be measured in terms of the resulting residual variance in the prediction of y_i.

2.3. SDC methods for tabular outputs

If a cell in a table is deemed sensitive, that is, the cell value represents an unacceptably high disclosure risk, a number of SDC approaches may be used.

2.3.1. Redefinition of cells

The cells are redefined to remove sensitive cells, for example, by combining sensitive cells with other cells or by combining categories of the cross-classified variables. This is also called *table redesign* (Willenborg and de Waal, 2001).

2.3.2. Cell suppression

The value of a sensitive cell is suppressed. Depending upon the nature of the table and its published margins, it may also be necessary to suppress the values of "complementary" cells to prevent an intruder being able to deduce the value of the cell from other values in the table. There is a large literature on approaches to choosing complementary cells which ensure disclosure protection. See, for example, Willenborg and de Waal (2001), Cox (2001), and Giessing (2001) and references therein.

2.3.3. Cell modification

The cell values may be modified in some way. It will generally be necessary to modify not only the values in the sensitive cells but also values in some complementary nonsensitive cells, for the same reason as in cell suppression. Modification may be deterministic, for example, Cox et al. (2004), or stochastic, for example, Willenborg and de Waal (2001, Section 9.2). A simple method is *rounding*, where the modified cell values are multiples of a given base integer (Willenborg and de Waal, 2001, Chapter 9). This method is more commonly applied to frequency tables derived from 100% data but can also be applied to tables of estimated totals from surveys, where the base integer may be chosen according to the magnitudes of the estimated totals. Instead of replacing the cell values by single safe values, it is also possible to replace the values by intervals, defined by lower and upper bounds (Salazar, 2003; Giessing and Dittrich, 2006). The method of *controlled tabular adjustment* (Cox et al., 2004) determines modified cell values within

such bounds so that the table remains additive and certain safety and statistical properties are met.

2.3.4. Pretabular microdata modification

Instead of modifying the cell values, the underlying microdata may be perturbed, for example, by adding noise, and then the table formed from the perturbed microdata (Evans et al., 1998; Massell et al., 2006).

The statistical output from a survey will typically include many tables. Although the above methods may be applied separately to each table, such an approach takes no account of the possible additional disclosure risks arising from the combination of information from different tables, in particular, from common margins. To protect against such additional risks raise new considerations for SDC. Moreover, the set of tables constituting the statistical output is not necessarily fixed, as in a traditional survey report. With developments in online dissemination, there is increasing demand for the generation of tables which can respond in a more flexible way to the needs of users. This implies the need to consider SDC methods which not only protect each table separately as above but also protect against the risk arising from alternative possible sequences of released tables (see, e.g., Dobra et al., 2003).

3. Microdata

3.1. Assessing disclosure risk

We suppose the agency is considering releasing to researchers an anonymized microdata file, where the records of the file correspond to the basic analysis units and each record contains a series of survey variables. The record may also include identifiers for higher level analysis units, for example, household identifiers where the basic units are individuals, as well as information required for survey analysis such as survey weights and primary sampling unit (PSU) identifiers.

We suppose that the threat of concern is that an intruder may link a record in the file to some external data source of known units using some variables, which are included in both the microdata file and the external source. These variables are often called *key variables* or identifying variables. There are various ways of defining disclosure risk in this setting. See, for example, Paass (1988) and Duncan and Lambert (1989). A common approach, often motivated by the nature of the confidentiality pledge, is to consider a form of *identification risk* (Bethlehem et al., 1990; Reiter, 2005), concerned with the possibility that the intruder will be able to determine a correct link between a microdata record and a known unit. This definition of risk will only be appropriate if the records in the microdata can meaningfully be said to be associated with units in the population. When microdata is subject to some forms of SDC, this may not be the case (e.g., if the released records are obtained by combining original records) and in this case, it may be more appropriate to consider some definition of predictive disclosure (e.g., Fuller, 1993) although we do not pursue this further here.

A number of approaches to the assessment of identification risk are possible, but all depend importantly upon assumptions about the nature of the key variables. One approach is to conduct an empirical experiment, matching the proposed microdata

against another data source, which is treated as a surrogate for the data source held by the intruder. Having made assumptions about the key variables, the agency can use record linkage methods (see Chapter 14), which it is plausible would be available to an intruder, to match units between the two data sets. Risk might then be measured in terms of the number of units for which matches are achieved together with a measure of the match quality (in terms of the proportions of false positives and negatives). Such an experiment, therefore, requires that the agency has information which enables it to establish precisely which units are in common between the two sources and which are not.

The key challenge in this approach is how to construct a realistic surrogate intruder data set, for which there is some overlap of units with the microdata and the nature of this overlap is known. On some occasions a suitable alternative data source may be available. Blien et al. (1992) provide one example of a data source listing people in certain occupations. Another possibility might be a different survey undertaken by the agency, although agencies often control samples to avoid such overlap. Even if there is overlap, say with a census, determining precisely which units are in common and which are not may be resource intensive. Thus, this approach is unlikely to be suitable for routine use.

In the absence of another data set, the agency may consider a reidentification experiment, in which the microdata file is matched against itself in a similar way, possibly after the application of some SDC method (Winkler, 2004). This approach has the advantage that it is not model-dependent, but it is possible that the reidentification risk is overestimated if the disclosure protection effects of sampling and measurement error are not allowed for in a realistic way.

In the remainder of Section 3, we consider a third approach, which again only requires data from the microdata file, but makes theoretical assumptions, especially of a modeling kind, to estimate identification risk. As for the reidentification experiment, this approach must make assumptions about how the key variables are measured in the microdata and by the intruder on known units using external information. A simplifying but "worst case" assumption is that the key variables are recorded in identical ways in the microdata and externally. We refer to this as the *no measurement error assumption*, since measurement error in either of the data sources may be expected to invalidate this assumption. If at least one of the key variables is continuous and the no measurement error assumption is made, then an intruder who observes an exact match between the values of the key variables in the microdata and on the known units could conclude with probability one that the match is correct, in other words, the identification risk would be one. If at least one of the key variables is continuous and it is supposed that measurement error may occur, then the risk will generally be below one. Moreover, an exact matching approach is not obviously sensible and a broader class of methods of record linkage might be considered. See Fuller (1993) for the assessment of disclosure risk under some measurement error model assumptions.

In practice, variables are rarely recorded in a continuous way in social survey microdata. For example, age would rarely be coded with more detail than 1 year bands. And from now on, we restrict attention to the case of categorical key variables. For simplicity, we restrict attention to the case of exact matching, although more general record linkage methods could be used. We focus on a microdata file, where the only SDC methods which have been applied are recoding of key variables or random

(sub)sampling. We comment briefly on the impact of other SDC methods on risk in Section 3.4.

3.2. *File-level measures of identification risk*

We consider a finite population U of N units (which will typically be individuals) and suppose the microdata file consists of records for a sample $s \subset U$ of size $n \leq N$. We assume that the possibility of statistical disclosure arises if an intruder gains access to the microdata and attempts to match a microdata record to external information on a known unit using the values of m categorical key variables $X_1, \ldots, X_m$. (Note that s and $X_1, \ldots, X_m$ are defined after the application of (sub)sampling or recoding, respectively, as SDC methods to the original microdata file.)

Let the variable formed by cross-classifying $X_1, \ldots, X_m$ be denoted by X, with values denoted $k = 1, \ldots, K$, where K is the number of categories or key values of X. Each of these key values corresponds to a possible combination of categories of the key variables. Under the no measurement error assumption, identity disclosure is of particular concern if a record is unique in the population with respect to the key variables. A record with key value k is said to be *population unique* if $F_k = 1$, where F_k denotes the number of units in U with key value k. If an intruder observes a match with a record with key value k, knows that the record is population unique and can make the no measurement error assumption then the intruder can infer that the match is correct.

As a simple measure of disclosure risk, we might therefore consider taking some summary of the extent of population uniqueness. In survey sampling, it is usual to define parameters of interest at the population level and this might lead us to define our measure as the population proportion N_1/N, where $N_r = \sum_k I(F_k = r)$ is the population frequencies of frequencies, $r = 1, 2, \ldots$. From a disclosure risk perspective, however, we are interested in the risk for a specific microdata file it is natural to allow the risk measure to be sample dependent. Thus, we might expect the risk to be higher if a sample is selected with a high proportion of unusual identifiable units than for a sample where this proportion is lower. Thus, a more natural file-level measure is the proportion of population uniques in the sample. Let the sample counterpart of F_k be denoted by f_k, then this measure can be expressed as follows:

$$\Pr(PU) = \sum_k I(f_k = 1, F_k = 1)/n. \tag{2}$$

It could be argued, however, that the denominator of this proportion should be made even smaller, since the only records which might possibly be population unique are ones that are sample unique (since $f_k \leq F_k$), that is, have a key value k such that $f_k = 1$. Thus, a more conservative measure would be to take

$$\Pr(PU|SU) = \sum_k I(f_k = 1, F_k = 1)/n_1, \tag{3}$$

where n_1 is the number of sample uniques and, more generally, $n_r = \sum_k I(f_k = r)$ is the sample frequencies of frequencies. For further consideration of the proportion of sample uniques that are population unique, see Fienberg and Makov (1998) and Samuels (1998).

It may be argued (e.g., Skinner and Elliot, 2002) that these measures may be overoptimistic, since they only capture the risk arising from population uniques and not from other records with $F_k \geq 2$. If an intruder observes a match on a key value with frequency F_k, then (subject to the no measurement error assumption) the probability that the match is correct is $1/F_k$ under the exchangeability assumption that the intruder is equally likely to have selected any of the F_k units in the population. An alternative measure of risk is then obtained by extending this notion of probability of correct match across different key values. Again, on worst case grounds, it is natural to restrict attention to sample uniques. One measure arises from supposing that the intruder starts with the microdata, is equally likely to select any sample unique and then matches this sample unique to the population. The probability that the resulting match is correct is then the simple average of $1/F_k$ across sample uniques:

$$\theta_s = \left[\sum_k I(f_k = 1)/F_k \right] / n_1. \tag{4}$$

Another measure is

$$\theta_U = \sum_k I(f_k = 1) \Big/ \sum_k F_k I(f_k = 1), \tag{5}$$

which is the probability of a correct match under a scenario where the intruder searches at random across the population and finds a match with a sample unique.

All the above four measures are functions of both the f_k and the F_k. The agency conducting the survey will be able to determine the sample quantities f_k from the microdata but the population quantities F_k will generally be unknown. It is, therefore, of interest to be able to make inference about the measures from sample data.

Skinner and Elliot (2002) showed that, under Bernoulli sampling with inclusion probability π, a simple design-unbiased estimator of θ_U is $\hat{\theta}_U = n_1/[n_1 + 2(\pi^{-1} - 1)n_2]$. They also provided a design consistent estimator for the asymptotic variance of $\hat{\theta}_U - \theta_U$. Skinner and Carter (2003) showed that a design-consistent estimator of θ_U for an arbitrary complex design is $\hat{\theta}_U = n_1/[n_1 + 2(\overline{\pi}_2^{-1} - 1)n_2]$, where $\overline{\pi}_2^{-1}$ is the mean of the inverse inclusion probabilities π_i^{-1} for units i with key values for which $f_k = 2$. They also provided a design-consistent estimator of the asymptotic variance of $\hat{\theta}_U - \theta_U$ under Poisson sampling.

Such simple design-based inference does not seem to be possible for the other three measures in (2)–(4). Assuming a symmetric design, such as Bernoulli sampling, we might suppose that $n_1, n_2, \ldots$ represent sufficient statistics and seek design-based moment-based estimators of the measures by solving the equations:

$$E(n_r) = \sum_t N_t P_{rt}, \quad r = 1, 2, \ldots,$$

where the coefficients P_{rt} are known for sampling schemes, such as simple random sampling or Bernoulli sampling (Goodman, 1949). The solution of these equations for N_t with $E(n_r)$ replaced by n_r gives unbiased estimators of K and N_1 under apparently weak conditions (Goodman, 1949). Unfortunately, Goodman found that the estimator of K can be "very unreasonable" and the same appears to be so for the corresponding estimator of N_1. Bunge and Fitzpatrick (1993) reviewed approaches to estimating K and

discussed these difficulties. Zayatz (1991) and Greenberg and Zayatz (1992) proposed an alternative "nonparametric" estimator of N_1 but this appears to be subject to serious upward bias for small sampling fractions (Chen and Keller-McNulty, 1998).

One way of addressing these estimation difficulties is by making stronger modeling assumptions, in particular by assuming that the F_k are independently distributed as follows:

$$F_k|\lambda_k \sim \text{Po}(\lambda_k) \tag{6}$$

where the λ_k are independently and identically distributed, that is, that the F_k follow a compound Poisson distribution. A tractable choice for the distribution of λ_k is the gamma distribution (Bethlehem et al., 1990) although it does not appear to fit well in some real data applications (e.g., Chen and Keller-McNulty, 1998; Skinner et al., 1994). A much better fit is provided by the log-normal (Skinner and Holmes, 1993). Samuels (1998) discussed estimation of $\Pr(PU|SU)$ based on a Poisson-Dirichlet model. A general conclusion seems to be that results can be somewhat sensitive to the choice of model, especially as the sampling fraction decreases, and that θ_U can be more robustly estimated than the other three measures.

3.3. Record-level measures of identification risk

A concern with file-level measures is that the principles governing confidentiality protection often seek to avoid the identification of *any* individual, that is require the risk to be below a threshold for each record, and such aims may not adequately be addressed by aggregate measures of the form (2)–(5). To address this concern, it is more natural to consider record level measures, that is, measures which may take different values for each microdata record. Such measures may help identify those parts of the sample where risk is high and more protection is needed and may be aggregated to a file level measure in different ways if desired (Lambert, 1993). Although record level measures may provide greater flexibility and insight when assessing whether specified forms of microdata output are "disclosive," they are potentially more difficult to estimate than file level measures.

A number of approaches have been proposed for the estimation of record level measures. For continuous key variables, Fuller (1993) showed how to assess the record level probability of identification in the presence of added noise, under normality assumptions. See also Paass (1988) and Duncan and Lambert (1989). We now consider related methods for categorical variables, following Skinner and Holmes (1998) and Elamir and Skinner (2006).

Consider a microdata record with key value X. Suppose the record is sample unique, that is, with a key value k for which $f_k = 1$, since such records may be expected to be most risky. Suppose the intruder observes an exact match between this record and a known unit in the population. We make the no measurement error assumption so that there will be F_k units in the population which potentially match the record. We also assume no response knowledge (see Section 2.1). The probability that this observed match is correct is

$$\Pr(\text{correct match} \mid \text{exact match}, X = k, F_k) = 1/F_k, \tag{7}$$

where the probability distribution is with respect to the design under a symmetric sampling scheme, such as simple random sampling or Bernoulli sampling. (Alternatively, it could be with respect to a stochastic mechanism used by the intruder, which selects any of the F_k units with equal probability). This probability is conditional on the key value k and on F_k.

In practice, we only observe the sample frequencies f_k and not the F_k. We, therefore, integrate out over the uncertainty about F_k and write the measure as

$$\Pr(\text{correct match} \mid \text{exact match}, X = k, f_k) = E(1/F_k|k, f_k = 1). \tag{8}$$

This expectation is with respect to both the sampling scheme and a model generating the F_k, such as the compound Poisson model in (6). An alternative measure, focusing on the risk from population uniqueness, is

$$\Pr(F_k = 1|k, f_k = 1). \tag{9}$$

The expressions in (8) and (9) may be generalized for any record in the microdata with $f_k > 1$. A difference between the probabilities in (8) and (9) and those in the previous section is that here we condition on the record's key value $X = k$. Thus, although we might assume $F_k|\lambda_k \sim \text{Po}(\lambda_k)$, as in (6), we should like to condition on the particular key value k when considering the distribution of λ_k. Otherwise, if the λ_k is identically distributed as in the previous section, then we would obtain the same measure of risk for all (sample unique) records. A natural model is a log-linear model:

$$\log(\lambda_k) = z_k\beta, \tag{10}$$

where z_k is a vector of indicator variables representing the main effects and the interactions between the key variables $X_1, \ldots, X_m$, and β is a vector of unknown parameters.

Expressions for the risk measures in (8) and (9) in terms of β are provided by Skinner and Holmes (1998) and Elamir and Skinner (2006). Assumptions about the sampling scheme are required to estimate β. Under Bernoulli sampling with inclusion probability π, it follows from (6) that $f_k|\lambda_k \sim \text{Po}(\pi\lambda_k)$. Assuming also (10), β may be estimated by standard maximum likelihood methods. A simple extension of this argument also applies under Poisson sampling where the inclusion probability π_k may vary with respect to the key variables, for example, if a stratifying variable is included among the key variables. In this case, we have $f_k|\lambda_k \sim \text{Po}(\pi_k\lambda_k)$. Skinner and Shlomo (2008) discussed methods for the specification of the model in (10). Skinner (2007) discussed the possible dependence of the measure on the search method used by the intruder.

3.4. SDC methods

In this section, we summarize a number of SDC methods for survey microdata.

3.4.1. Transformation of variables to reduce detail

Categorical key variables may be transformed, in particular, by combining categories. For example, the variable household size might be *top coded* by creating a single maximum category, such as 8+. Continuous key variables may be *banded* to form ordinal categorical variables by specifying a series of cut-points between which the intervals define categories. The protection provided by combining categories of key variables

can be assessed following the methods in Sections 3.2 and 3.3. See also Reiter (2005). Provided the transformation is clear and explicit, this SDC method has the advantage that the reduction of utility is clear to the data user, who may suffer loss of information but the validity of analyses is not damaged.

3.4.2. Stochastic perturbation of variables

The values of potential key variables are perturbed in a stochastic way. In the case of continuous variables, perturbation might involve the *addition of noise*, analogous to the addition of measurement error (Fuller, 1993; Sullivan and Fuller, 1989). In the case of categorical variables, perturbation may consist of misclassification, termed the *Postrandomization Method* (PRAM) by Gouweleeuw et al. (1998). Perturbation may be undertaken in a way to preserve specified features of the microdata, for example, the means and standard deviations of variables in the perturbed microdata may be the same as in the original microdata, but in practice there will inevitably be unspecified features of the microdata which are not reproduced. For example, the estimated correlation between a perturbed variable and an unperturbed variable will often be downwardly biased if an analyst uses the perturbed data but ignores the fact that perturbation has taken place. An alternative is to provide users with the precise details of the perturbation method, including parameter values, such as the standard deviation of the noise or the entries in the misclassification matrix, so that they may "undo" the impact of perturbation when undertaking their analyses. See, for example, Van den Hout and Van der Heijden (2002) in the case of PRAM or Fuller (1993) in the case of added noise. In principle, this may permit valid analyses although there will usually be a loss of precision and the practical disadvantages are significant.

3.4.3. Synthetic microdata

This approach is similar to the previous approach, except that the aim is to avoid requiring special methods of analysis. Instead, the values of variables in the file are replaced by values generated from a model in a way that is designed for the analysis of the synthetic data, as if it were the true data, to generate consistent point estimates (under the assumption that the model is valid). The model is obtained from fitting to the original microdata. To enable valid standard errors as well as consistent point estimators, Raghunathan et al. (2003) proposed that multiple copies of the synthetic microdata are generated in such a way that multiple imputation methodology can be used. See Reiter (2002) for discussion of complex designs. Abowd and Lane (2004) discussed release strategies combining remote access to one or more such synthetic microdata files with much more restricted access to the original microdata in a safe setting.

3.4.4. Selective perturbation

Often concern focuses only on records deemed to be risky and it may be expected that utility will be greater if only a subset of risk records is perturbed. In addition to creating stochastically perturbed or synthetic values for only targeted records, it is also possible just to create missing values in these records, called *local suppression* by Willenborg and de Waal (2001), or both to create missing values and to replace these by imputed values, called *blank and impute* by Federal Committee on Statistical Methodology (2005). A major problem with such methods is that they are likely to create biases if the targeted values are unusual. The data user will typically not be able

to quantify these biases, especially when the records selected for blanking depend on the values of the variable(s) which are to made missing. Reiter (2003) discussed how valid inference may be conducted if multiple imputed values are generated in a specified way for the selected records. He referred to the resulting data as *partially synthetic microdata*.

3.4.5. Record swapping

The previous methods focus on the perturbation of the values of the variables for all or a subset of records. The method of record swapping involves, instead, the values of one or more key variables being swapped between records. The choice of records between which values are swapped may be controlled so that certain bivariate or multivariate frequencies are maintained (Dalenius and Reiss, 1982) in particular by only swapping records sharing certain characteristics (Willenborg and de Waal, 2001, Section 5.6). In general, however, it will not be possible to control all multivariate relationships and record swapping may damage utility in an analogous way to misclassification (Skinner and Shlomo, 2007). Reiter (2005) discussed the impact of swapping on identification risk.

3.4.6. Microaggregation

This method (Defays and Anwar, 1998) is relevant for continuous variables, such as in business survey microdata, and in its basic form consists of ordering the values of each variable and forming groups of a specified size k (the first group contains the k smallest values, the second group the next k smallest values, and so on). The method replaces the values by their group means, separately for each variable. An advantage of the method is that the modification to the data will usually be greatest for outlying values, which might also be deemed the most risky. It is difficult, however, for the user to assess the biasing impact of the method on analyses.

SDC methods will generally be applied after the editing phase of the survey, during which data may be modified to meet certain edit constraints (see Chapter 9). The application of some SDC methods may, however, lead to failure of some of these constraints. Shlomo and de Waal (2006) discussed how SDC methods may be adapted to take account of editing considerations.

3.5. SDC for survey weights and other design information

Survey weights and other complex design information are often released with survey microdata in order that valid analyses can be undertaken. It is possible, however, that such design information may contribute to disclosure risk. For example, suppose a survey is stratified by a categorical variable X with different sampling fractions in different categories of X. Then, if the nature of the sampling design is published (as is common), it may be possible for the intruder to determine the categories of X from the survey weight. Thus, the survey design variable may effectively become a key variable. See de Waal and Willenborg (1997) and Willenborg and de Waal (2001, Section 5.7) for further discussion of how survey weights may lead to design variables becoming key variables. Note that this does not imply that survey weights should not be released; it just means that disclosure risk assessments should take account of what information survey weights may convey. Willenborg and de Waal (2001, Section 5.7.3) and Mitra and Reiter (2006) proposed some approaches to adjusting weights to reduce risk.

In addition to the release of survey weights, it is common to release either stratum or PSU labels or replicate labels, to enable variances to be estimated. These labels will generally be arbitrary and will not, in themselves, convey any identifying information. Nevertheless, as for survey weights, the possibility that they could be used to convey information indirectly needs to be considered. For example, if the PSUs are defined by areas for which public information is available, for example, a property tax rate, and the microdata file includes area-level variables, then it is possible that these variables may enable a PSU to be linked to a known area. As another example, suppose that a PSU is an institution, such as a school, then school level variables on the microdata file, such as the school enrolment size, might enable the PSU to be linked to a known institution. Even for individual level microdata variables, it is possible that sample-based estimates of the total or mean of such variables for a stratum, say, could be matched to published values, allowing for sampling uncertainty.

A standard simple approach to avoiding releasing PSU or replicate identifiers is to provide information on design effects or generalized variance functions instead. Such methods are often inadequate, however, for the full range of uses of survey microdata (Yung, 1997). Some possible more sophisticated approaches include the use of adjusted bootstrap replicate weights (Yung, 1997), adjusted pseudoreplicates or pseudo PSU identifiers (Dohrmann et al., 2002), or combined stratum variance estimators (Lu et al., 2006).

4. Conclusion

The development of SDC methodology continues to be stimulated by a wide range of practical challenges and by ongoing innovations in the ways that survey data are used, with no signs of diminishing concerns about confidentiality. There has been a tendency for some SDC methods to be developed in somewhat ad hoc way to address specific problems, and one aim of this chapter has been to draw out some principles and general approaches which can guide a more unified methodological development. Statistical modeling has provided one important framework for this purpose. Other fields with the potential to influence the systematic development of SDC methodology in the future include data mining, in particular methods related to record linkage and approaches to privacy protection in computer science and database technology.

Acknowledgments

I am grateful to Natalie Shlomo and a reviewer for comments on an earlier draft.