Journal of Official Statistics
Vol. 9. No. 2, 1993, pp. 407–426

Statistical Analysis of Masked Data

Roderick J.A. Little[1]

Abstract: A model-based likelihood theory is presented for the analysis of data masked for confidentiality purposes. The theory builds on frameworks for missing data and treatment assignment, and a theory for coarsened data. It distinguishes a model for the masking selection mechanism, which determines which data values are masked, and the masking treatment mechanism, which specifies how the masking is carried out. The framework is applied to a variety of masking methods, including randomized response, subsampling of cases or variables, deletion, coarsening by grouping or rounding, imputation, aggregation, noise injection and simulation of artificial records.

Key words: Confidentiality; grouped data; imputation; missing data; randomized response; rounded data; slicing.

1. Introduction

1.1. The problem

Increased concern for protecting the confidentiality of respondents of censuses and surveys is evidenced by extensive recent interest and research in masking methods. In the future, users of public use files and other products of statistical agencies will be faced increasingly with files that have been altered to protect the privacy of respondents. The major focus of research in this area has been on definitions and measures of disclosure risk (e.g., Duncan and Lambert 1986, 1989; Paass 1988; Bethlehem, Keller and Pannekoek 1990), the development and choice of masking methods (e.g., Kim 1986; McGuckin and Nguyen 1990; Sullivan and Fuller, 1989, 1990; Greenberg 1990), and the ability of methods to enhance confidentiality (e.g., Paass 1988). This paper discusses masking from the point of view of statistical analysis of the resulting data. A general likelihood-based framework is developed for masking and analysis of microdata files, and applied in a variety of masking settings. Analysis issues for specific masking procedures are discussed, and areas of future research on the analysis of masked data are identified.

Masking methods can be applied (a) when the data are collected, using methods such as randomized response, (b) when the data are supplied to the user for analysis, for example, by deleting or altering values in a public use tape, or (c) when the results of an analysis are presented, for example, by deleting cell counts in a cross-tabulation. In this article I shall be primarily concerned with (a) and (b), although some of my remarks have implications for (c). A number of masking

[1] Department of Biomathematics, UCLA School of Medicine, Los Angeles, CA 90024-1766, U.S.A.

methods will be discussed here, including randomized response; release of subsamples of records; suppression of cells in a crosstabulation; deletion of sensitive values; deletion followed by imputation of values; addition of random noise; rounding, grouping or truncation; transformation; slicing files into subsets of variables; slicing and recombination to form synthetic records; reduction to aggregate sufficient statistics; simulation of artificial records; and microaggregation. Before considering analysis issues, some general thoughts on the nature of masking are offered.

1.2. *Aggregate analysis, individual protection*

Many have noted that masking is a double-edged sword, in that increased protection goes hand in hand with loss of information for analysis. It seems useful to focus not only on complementary properties of masking and analysis, but also on properties that distinguish the two activities. The key distinction is that masking is primarily concerned with identification of *individual* records, whereas statistical analysis is concerned with making inferences about *aggregates*. (The conception of aggregate analysis is broad here, including analytical methods such as regression as well as descriptive summaries such as means and totals.) Methods that exploit this distinction can achieve great gains in confidentiality at little cost. As a trivial example, rounding date of birth to year of birth may dramatically decrease the incidence of uniquely identifiable records in a file, with minor implications for many aggregate statistical analyses. On the other hand, masking becomes inherently difficult when the distinction between aggregate and individual is not clear-cut, as when one large firm dominates a business file, or analysis is required for small subdomains of the population.

Consider the following basic scenario for assessing masking methods. A data snooper attempts to identify respondents in a public use file to obtain additional information about them. The file includes *key* variables, which are known for one or more individuals by the snooper and can be used to identify a record, that is, establish a one-to-one correspondence between a record and a specific individual. Obvious key variables are name and address, but variables such as household composition, age, race, and occupation can also serve as keys (Bethlehem et al. 1990). The file also contains *target* variables that provide new information on identified individuals; these may be sensitive variables such as sexual activity or HIV status in a survey on AIDS. As Bethlehem et al. (1990) point out, a variable such as income may be sensitive in some cultures and hence considered a target, but less sensitive and more widely known in other cultures, where it may be classified as a key.

Methods that mask the key variables impede identification of the respondent in the file, and methods that mask the target variables limit what is learned if a match is made. Both approaches may be useful, and in practice a precise classification of variables as keys or targets may be difficult. However, masking of targets is more vulnerable to the trade-off between protection gain and information loss than masking of keys; hence masking of keys seems potentially more fruitful. For example, rounding date of birth as a key variable may serve to impede identification of individual respondents, but rounding date of birth as a target achieves little since the loss of information is minor.

Another illustration of asymmetry

between protection and information loss is that certain masking methods afford protection with no loss of information at all for certain analyses. For example, a file containing the mean and covariance matrix of a set of variables allows analyses based on these statistics (such as regression) to be carried out with full efficiency. Direct presentation of aggregate statistics seems a simple and powerful masking technique. The trade-off is not between privacy and information loss but between privacy and *flexibility* of analysis, since the data producer's choice of variables greatly restricts subsequent analysis. Methods that attempt to extend flexibility while maintaining protection, such as microaggregation or noise injection, seem worth pursuing. Methodology can help, but close communication between the analyst and the data collector seems essential to limit the effect of masking methods on flexibility of analysis.

2. A Likelihood Theory for Masked Data Files

2.1. *Formal theory*

Three key issues arise in masking data files for confidentiality:

a. *Selection*: which values in the data set should be masked?

b. *Treatment*: how should the values be masked?

c. *Analysis*: how should the resulting masked data be analyzed?

These three aspects can be formalized concisely within a model-based (likelihood) analysis perspective. The following theory combines elements of Rubin's (1976, 1978a) theories for treatment assignment and missing data, and Heitjan and Rubin's (1991) theory for coarsened data.

Let $\mathbf{X} = \{x_{ij}\}$ denote an $(n \times p)$ data matrix of n observations on p variables prior to masking, and $\mathbf{M} = \{m_{ij}\}$ denote the masking indicator matrix, with $m_{ij} = 1$ if x_{ij} is masked and $m_{ij} = 0$ otherwise. Let $\mathbf{Z} = \{z_{ij}\}$, where z_{ij} represents the masked value of x_{ij}; it is convenient to define z_{ij} for data that are not masked as well as for values that are masked. For deleted data z_{ij} would be a missing-value code; other examples are an imputed value or a recode representing a grouped version of x_{ij}. We model the joint distribution of $\mathbf{X}$, $\mathbf{Z}$ and $\mathbf{M}$ with density function

$$f(\mathbf{X}, \mathbf{Z}, \mathbf{M}|\boldsymbol{\theta}) = f_X(\mathbf{X}|\boldsymbol{\theta}) f_Z(\mathbf{Z}|\mathbf{X}) f_M(\mathbf{M}|\mathbf{X}, \mathbf{Z}). \quad (1)$$

Here $f_X(\mathbf{X}|\boldsymbol{\theta})$ is the density for the unmasked data with unknown parameters $\boldsymbol{\theta}$, which would form the basis for analysis in the absence of masking; $f_Z(\mathbf{Z}|\mathbf{X})$ is the distribution of the masked data values, which formalizes the masking treatment; and $f_M(\mathbf{M}|\mathbf{X}, \mathbf{Z})$ represents the distribution of the masking selection mechanism, which formalizes the selection of values that are masked. If the analyst knows which values are masked and the method of masking, then both $\mathbf{M}$ and the distributions of $\mathbf{Z}$ and $\mathbf{M}$ are known. If the analyst does not know which values are masked and which are not masked, then $\mathbf{M}$ is unknown. In other settings it may be necessary to index the distributions of $\mathbf{Z}$ and/or $\mathbf{M}$ by unknown parameters; for example, the data may be subjected to an unknown transformation (McGuckin and Nguyen 1990), or noise with unknown variance added. A full likelihood analysis would then involve both $\boldsymbol{\theta}$ and unknown masking parameters, with possible problems of parameter identification.

Now write $\mathbf{X} = (\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{mis}})$, $\mathbf{Z} = (\mathbf{Z}_{\text{obs}}, \mathbf{Z}_{\text{mis}})$, where obs denotes observed components and mis missing components of each matrix. Analysis of the masked data should be based on the likelihood for $\boldsymbol{\theta}$

given the data $\mathbf{M}$, $\mathbf{X}_{obs}$ and $\mathbf{Z}_{obs}$, which is obtained formally by integrating the joint density of $\mathbf{X}$, $\mathbf{Z}$ and $\mathbf{M}$ over the missing values ($\mathbf{X}_{mis}$ and $\mathbf{Z}_{mis}$) :

$$L(\boldsymbol{\theta}|\mathbf{M}, \mathbf{X}_{obs}, \mathbf{Z}_{obs}) = \int f_X(\mathbf{X}|\boldsymbol{\theta}) f_Z(\mathbf{Z}|\mathbf{X}) \times f_M(\mathbf{M}|\mathbf{X}, \mathbf{Z}) d\mathbf{X}_{mis} d\mathbf{Z}_{mis}.$$

The distribution of $\mathbf{M}$ in this expression may depend on $\mathbf{X}$ and possibly on $\mathbf{Z}_{obs}$, but should not depend on $\mathbf{Z}_{mis}$. Hence $f_M(\mathbf{M}|\mathbf{X}, \mathbf{Z}) = f_M(\mathbf{M}|\mathbf{X}, \mathbf{Z}_{obs})$, and integrating over $\mathbf{Z}_{mis}$ yields

$$L(\boldsymbol{\theta}|\mathbf{M}, \mathbf{X}_{obs}, \mathbf{Z}_{obs}) = \int f_X(\mathbf{X}|\boldsymbol{\theta}) f_Z^*(\mathbf{Z}_{obs}|\mathbf{X}) \times f_M(\mathbf{M}|\mathbf{X}, \mathbf{Z}_{obs}) d\mathbf{X}_{mis} \quad (2)$$

where $f_Z^*(\mathbf{Z}_{obs}|\mathbf{X}) = \int f_Z(\mathbf{Z}|\mathbf{X}) d\mathbf{Z}_{mis}$.

Applying the ideas of Rubin (1976, 1978a), if masking selection and treatment satisfy certain ignorability conditions then corresponding terms in the likelihood (2) can be omitted. Specifically, the masking selection mechanism is called *ignorable* if its distribution depends only on observed values in the masked data set, that is

$$f_M(\mathbf{M}|\mathbf{X}, \mathbf{Z}) = f_M(\mathbf{M}|\mathbf{X}_{obs}, \mathbf{Z}_{obs}) \quad \text{for all } \mathbf{X}_{mis}, \mathbf{Z}_{mis}. \quad (3)$$

The masking treatment mechanism is called ignorable if the distribution of masked values depends only on observed values of $\mathbf{X}$, that is

$$f_Z^*(\mathbf{Z}_{obs}|\mathbf{X}) = f_Z^*(\mathbf{Z}_{obs}|\mathbf{X}_{obs}) \text{ for all } \mathbf{X}_{mis}. \quad (4)$$

It is easy to show that if the masking selection mechanism is ignorable, then the density of $\mathbf{M}$ can be omitted from the likelihood (2), yielding the simpler form

$$L_1(\boldsymbol{\theta}) = \int f_X(\mathbf{X}|\boldsymbol{\theta}) f_Z^*(\mathbf{Z}_{obs}|\mathbf{X}) \, d\mathbf{X}_{mis}. \quad (5)$$

Similarly if the masking treatment mechanism is ignorable, the density of $\mathbf{Z}_{obs}$ can be omitted from (2), yielding the expression

$$L_2(\boldsymbol{\theta}) = \int f_X(\mathbf{X}|\boldsymbol{\theta}) f_M(\mathbf{M}|\mathbf{X}, \mathbf{Z}_{obs}|\mathbf{X}) \, d\mathbf{X}_{mis}. \quad (6)$$

Finally if both the selection and treatment mechanisms are ignorable, the likelihood is simply

$$L_3(\boldsymbol{\theta}) = \int f_X(\mathbf{X}|\boldsymbol{\theta}) \, d\mathbf{X}_{mis} \quad (7)$$

which is proportional to the marginal density of $\mathbf{X}_{obs}$.

The size and complexity of public use files from large surveys are difficult enough to analyze without the additional problems associated with masking, and analysis based on (2), (5), (6) or (7) may be considerably more complex than analysis of the original unmasked data, perhaps involving iterative algorithms not available in standard software packages. Hence masking methods that yield simple likelihoods are attractive. One strategy that can be helpful is to treat the unobserved values of $\mathbf{X}$ as missing-data, and apply tools for missing-data analysis. Two such tools, the EM algorithm and multiple imputation, are described in the Appendix ; other references are Dempster, Laird and Rubin (1977); Rubin (1987); and Little and Rubin (1987, 1989).

Approximate analysis methods might also be considered. A simple approach is to treat the masked values as the truth, that is, ignore the process of masking entirely. If masked values are not identified, this may be the only realistic analysis option. In the current setting this corresponds to basing inference on the "pseudo-likelihood" function

$$L^*(\boldsymbol{\theta}) = f_X(\mathbf{Y}|\boldsymbol{\theta}) \quad (8)$$

where $\mathbf{Y} = \{y_{ij}\}$ is obtained by treating the

masked values as truth:

$$y_{ij} = \begin{cases} x_{ij}, \text{ if } m_{ij} = 0 \\ z_{ij}, \text{ if } m_{ij} = 1 \end{cases}.$$

This approach is at best approximate, and the substitution of masked for true values needs to make sense; for example, it would not be appropriate if z_{ij} was a categorical recode representing a grouping of an underlying continuous x_{ij}. An interesting property of a masking method might be how closely an analysis based on (8) approximates a more precise analysis based on the correct likelihood. Also the development of simple approximate methods that improve on the analysis based on (8), such as Sheppard's (1898) well-known corrections for grouped data, appears worthwhile.

The next two sections apply this theory to a variety of masking problems, Section 3 concentrating on masking selection and Section 4 on the masking treatment.

3. Examples of Masking Selection Mechanisms

Masking is rarely applied to the entire data matrix. In this section we discuss strategies that selectively mask rows of the matrix, columns of the matrix, and combinations of the two.

Example 1. Random Subsampling of Rows (Cases)

Suppose entire cases are either masked or not masked. For case i, let $\mathbf{x}_i$ and $\mathbf{z}_i$ denote the vectors of true and masked values, and let $m_i = 1$ if row i is masked, $m_i = 0$ otherwise. A simple approach is to release a random sample of the cases, as is done for certain census products; the inclusion of a small fraction of the original data clearly reduces the chance of identifying particular respondents. Random subsampling is clearly an ignorable masking selection mechanism; formally $f_M(m_i = 1|\mathbf{x}_i, \mathbf{z}_i)$ is a constant that does not depend on $\mathbf{X}$ or $\mathbf{Z}$. An advantage is that analysis of the subsample is straightforward. The fraction of retained cases needs to be small to provide significant protection, so the deletion of the masked cases involves a severe information loss. Nevertheless, the method has uses for censuses, administrative record systems or very large surveys.

Example 2. Masking of Selected Cases

It is tempting to delete or mask cases that have high risk of identification. In surveys of businesses, these are often large companies. To represent selection of this type, let x_{i1} be the value of a variable X_1 measuring the size of a firm, and suppose

$$f_M(m_i = 1|\mathbf{x}_i, \mathbf{z}_i) = \pi(x_{1i}) \qquad (9)$$

a monotonically increasing function of x_{i1}. A special case is right censoring, where $\pi(x_{i1}) = 1$ if $x_{i1} > c$, and zero otherwise, but other forms of π may also be useful. The mechanism (9) is nonignorable since it depends on the value x_{i1} which is not observed for masked cases. Analyses that assume the mechanism as ignorable, for example, by treating the unmasked data as a random sample, are generally inappropriate. If a selection mechanism of the form (9) is contemplated and X_1 is not itself a target, then the values of X_1 might be retained for both masked and unmasked cases, with sensitive variables deleted for masked cases. The variable X_1 can be used as a covariate to adjust for selection on size of firm.

Example 3. Masking of Selected Columns (Variables)

Large gains of statistical efficiency are possible by restricting masking to particular variables; hence the choice of which vari-

ables are masked needs careful attention. If values of a particular variable are either all masked or all unmasked, then the masking mechanism does not depend on the data and hence is ignorable.

As noted in Section 1, it may be possible to distinguish between key variables, which are used to identify respondents, and target variables, which represent additional information that is learned after identification. It is then useful to distinguish between approaches that mask key variables and approaches that mask target variables. The argument in Section 1.2 suggests that masking of keys may be preferable to the extent that protection can be afforded with relatively minor information loss (for example, by rounding a variable to remove uniquenesses). On the other hand, sometimes masking is confined to a particularly sensitive target variable, as in the randomized response technique discussed in Section 4.2.1.

Example 4. Masking of Selected Rows and Columns

Further efficiency gains are possible by restrictively masking a subset of variables for a subset of cases. If the masking method is deletion of a set of variables, the masked data then have the structure of a double sample, with unmasked variables measured for the whole sample, and masked variables available for a subsample of cases. The selection mechanism is ignorable if selection depends on the values of unmasked variables but not on the values of masked variables; the unmasked variables can be used as design variables for the selection of cases subject to masking. The masking selection design might be considered explicitly as an aspect of the survey design in surveys where confidentiality issues are important.

4. Masking Methods

4.1. Introduction

The following masking methods are prevalent in the literature:

a. *Deletion*, or more precisely, replacing the observed value by a missing-value code.
b. *Coarsening*, that is mapping particular values to a set of values. Continuous variables might be mapped into an interval, categorical variables recoded into a smaller set of variables.
c. *Imputation*, that is replacement of the true value by a substitute. A number of variants are discussed below.
d. *Aggregation*, that is presentation of data in aggregate form.

Generally speaking, deletion and coarsening place the added analysis burden of masking on the user. Imputation creates a rectangular file that is more amenable to analysis, but modifications to the complete-data analysis may be required to allow for imputation error. Aggregation limits the flexibility of the analysis. I now discuss these methods in more detail.

4.2. Masking by deletion, with or without imputation

4.2.1. Randomized response

Most masking procedures are applied after the data are collected. In contrast, randomized response (Warner 1965) masks the true response to a sensitive question at the time the question is asked. An obvious advantage is that the respondent actively participates in the masking, rather than having to rely on later actions by the data collector. Disadvantages include the loss of information entailed by randomizing the response, the added complexity of the inter-

viewing process, and the fact that the true data are not available to the data collector, where they might be made available to the analyst in a variety of masked forms depending on the nature of the file.

Randomized response data can be regarded as masked data with a deleted binary key variable Q which indicates whether the sensitive or control question was answered. One approach to analysis is to treat the values of Q as missing data and apply missing-data techniques such as the EM algorithm and multiple imputation. A detailed discussion for the many variants of randomized response is not attempted here; instead ideas are sketched for one form of the method, involving an unrelated question with known outcome probability.

Example 5. Randomized Response with an Unrelated Question.

Suppose the respondent is directed to answer one of the following two questions, depending on the outcome (known only to the respondent) of a randomizing procedure such as throwing a fair die;

1. The sensitive question of interest; for example, "Have you tested positive for the HIV virus?"; or

0. An unrelated question with known constant outcome probability μ; for example, "Were you born in May or November?" which might be regarded as having probability $\mu = 1/6$ to an acceptable degree of approximation.

Let $Q = j$ if question j was answered ($j = 1, 0$), and suppose

$$p(Q = 1) = 1 - p(Q = 0) = \pi$$

the known probability of receiving the sensitive question. Let Z denote the outcome of the randomized question, and suppose that

$$p(Z = 1|Q = 0) = \mu;$$

$$p(Z = 1|Q = 1) = \lambda(\mathbf{v})$$

where $\lambda(\mathbf{v})$ is the unknown probability associated with the sensitive question (for example, the probability of reporting HIV positive), which depends on a vector $\mathbf{v}$ of directly observed characteristics of the respondent (age, education, etc.). If Q is observed for every respondent, inferences about $\lambda(\mathbf{v})$ can be obtained directly by regressing the binary outcome Z on $\mathbf{V}$ using only cases with $Q = 1$. For example, one might apply logistic regression based on the model

$$\log[\lambda(\mathbf{v};\boldsymbol{\theta})/\{1 - \lambda(\mathbf{v};\boldsymbol{\theta})\}] = \boldsymbol{\theta}^{\mathrm{T}}\mathbf{v} \qquad (10)$$

where $\boldsymbol{\theta}$ is a vector of unknown regression coefficients. The problem is then to imitate this analysis when the value of Q is unobserved for all respondents.

The loglikelihood of $\boldsymbol{\theta}$ given a simple random sample $\mathbf{Y} = \{z_i, \mathbf{v}_i : i = 1, ..., n\}$ has the form

$$l(\boldsymbol{\theta}|\mathbf{Y}) = \sum_{i:z_i=1} \log\{\pi\lambda(\mathbf{v}_i;\boldsymbol{\theta}) + (1 - \pi)\mu\} + \sum_{i:z_i=1} \log\{1 - \pi\lambda(\mathbf{v}_i;\boldsymbol{\theta}) - (1 - \pi)\mu\} \qquad (11)$$

which can be maximized directly using a scoring algorithm. Alternatively, we can apply the EM algorithm, treating the values of Q as missing covariates. Ibrahim (1990) shows that for a broad class of models with incomplete categorical regressors, EM reduces to the following iteratively reweighted algorithm: given current parameters $\boldsymbol{\theta}^{(t)}$, the E-step computes the weight

$$w_i^{(t)} = E(Q_i|z_i, \mathbf{v}_i, \boldsymbol{\theta}^{(t)})$$

$$= p\{Q_i = 1|z_i, \mathbf{v}_i, \boldsymbol{\theta}^{(t)}\}$$

for each case i; the M-step computes $\boldsymbol{\theta}^{(t+1)}$ by standard complete-data ML, with case i weighted by $w_i^{(t)}$; for the model (11) this corresponds to weighted logistic regression. Successive iterates $\{\boldsymbol{\theta}^{(t)}\}$ converge to the ML estimate for the likelihood (11). A simple application of Bayes' Theorem yields

$$w_i^{(t)} = \frac{\pi\lambda(\mathbf{v}_i, \boldsymbol{\theta}^{(t)})}{(1-\pi)\mu + \pi\lambda(\mathbf{v}_i, \boldsymbol{\theta}^{(t)})} \tag{12}$$

if $z_i = 1$, and

$$w_i^{(t)} = \frac{\pi\{1-\lambda(\mathbf{v}_i, \boldsymbol{\theta}^{(t)})\}}{(1-\pi)(1-\mu) + \pi\{1-\lambda(\mathbf{v}_i, \boldsymbol{\theta}^{(t)})\}} \tag{13}$$

if $z_i = 0$. These weights measure both the loss of information and the gain of privacy from masking; small weights indicate poor ability to predict which question was answered, but also imply low analytical efficiency; and vice versa. Bourke and Moran (1988) give another application of EM to randomized response data.

An alternative analysis approach is to multiply-impute the question indicator Q, using draws from its predictive distribution. A valid asymptotic analysis proceeds as follows: for the mth set of imputations, a value $\boldsymbol{\theta}_{(m)}$ of $\boldsymbol{\theta}$ is drawn from a multivariate normal distribution centered at the ML estimate $\hat{\boldsymbol{\theta}}$ with covariance matrix given by the information matrix. Alternatively, $\boldsymbol{\theta}_{(m)}$ can be computed as the ML estimate of $\boldsymbol{\theta}$ for a bootstrap sample of cases, as in Heitjan and Little (1991). Then for each observation i, (a) the probability w_{im} is computed from equations (12) or (13), with $\boldsymbol{\theta}_{(m)}$ substituted for $\boldsymbol{\theta}^{(t)}$; and (b) Q_{im} is computed as a Bernoulli draw with probability w_{im}.

4.2.2. Slicing

Suppose the data set includes two blocks of variables $\mathbf{X}_1$ and $\mathbf{X}_2$, such that knowledge of the values of $\mathbf{X}_1$ and $\mathbf{X}_2$ may identify individuals in the file that cannot be identified from values $\mathbf{X}_1$ or $\mathbf{X}_2$ alone. One approach to masking is to provide separate files containing data on $\mathbf{X}_1$ and data on $\mathbf{X}_2$, but to omit information that allows individual values of $\mathbf{X}_1$ and $\mathbf{X}_2$ to be linked. This operation is called *slicing*.

Slicing can be couched as a form of variable deletion by introducing a permutation variable P that links the two files; that is, $P(i) = j$ if the ith case in the $\mathbf{X}_1$-file maps into the jth case in the $\mathbf{X}_2$-file. The unmasked data then consist of $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \mathbf{P})$, where rows of $\mathbf{X}_1$ and $\mathbf{X}_2$ are linked by $\mathbf{P}$. Slicing is then equivalent to deleting the values of $\mathbf{P}$.

The likelihood prior to masking is proportional to $f_X(\mathbf{X}_1, \mathbf{X}_2|\mathbf{P}, \boldsymbol{\theta})$, the density given the true value of $\mathbf{P}$. The likelihood of the masked data is proportional to

$$L(\boldsymbol{\theta}|\mathbf{Z}) = \sum_{\mathbf{P}} | f_X(\mathbf{X}_1, \mathbf{X}_2|\mathbf{P}, \boldsymbol{\theta}) \tag{14}$$

where the sum is over all possible permutations $\mathbf{P}$ linking the two files.

The marginal distributions of $\mathbf{X}_1$ and $\mathbf{X}_2$ are unaffected by the permutation $\mathbf{P}$, so analyses involving only $\mathbf{X}_1$ or $\mathbf{X}_2$ are clearly easy. Exact likelihood analysis involving data on both $\mathbf{X}_1$ and $\mathbf{X}_2$ should be based on (14), and is complicated by the extremely large number of permutations involved in the summation: even evaluating this likelihood is impractical except for very small files. Hence approximate analysis methods are needed. One such approach is for the data producer to provide an imputed permutation $\hat{\mathbf{P}}$ from the set of permutations that are relatively likely given the values of $\mathbf{X}_1$ and $\mathbf{X}_2$. Then inference is based on $f_X(\mathbf{X}_1, \mathbf{X}_2|\hat{\mathbf{P}}, \boldsymbol{\theta})$. Replacing $\mathbf{P}$ by $\hat{\mathbf{P}}$ has the effect of switching data between cases. If $\hat{\mathbf{P}}$ and $\mathbf{P}$ are chosen to be the same except on

Table 1. Hypothetical census data on two key variables at a particular location

		Size of household				
		1	2–3	4–6	>6	all
Race	1	116	284	85	10	495
	2	44	161	68	6	279
	3	11	23	3	1	38
	all	171	468	156	17	812

Location $(L) = 1$.

Table 2. Data from Table 1 further classified by a target variable (income)

		Size of household				
		1	2–3	4–6	>6	all
Income	< 30K	17	48	7	3	75
	30–60K	18	73	50	2	143
	> 60K	9	40	11	1	61
	all	44	161	68	6	279

Location $(L) = 1$, Race $= 2$.

a subset of cases, the effect is to limit slicing and switching to a subset of records. A good choice of $\hat{\mathbf{P}}$ provides confidentiality while yielding an approximate likelihood that is close to the likelihood given the true permutation $\mathbf{P}$.

A limitation of this approach is the failure to allow for the added uncertainty of imputation of $\mathbf{P}$. One possible extension would be to multiply-impute a set of permutations, thus allowing the propagation of error from imputing $\mathbf{P}$. Ideally the permutations should be drawn from the posterior distribution of $\mathbf{P}$, which involves considerable computation; however, simpler approaches might provide a step in the right direction.

4.2.3. Masking data with a set of categorical key variables

An important masking problem arises when cases can be uniquely identified in a file from a set of K key variables (Bethlehem et al. 1990). If the key variables are categorical, data on them can be arranged in a K-way contingency table, and uniquenesses are single counts in this table; the masking problem is then to avoid disclosure of target variable values for these cases; the disclosure risk may also be considered high in cells with a small number of cases (say between two and nine), and cases in these "sensitive cells" might also be masked.

Example 6. A Numerical Example with Three Categorical Keys

Consider for concreteness a three-way table of counts of households from a census, classified by three key variables, S = household size (4 categories), R = Race (3 categories), and L = Location; hypothetical data for 812 households in one location are presented in Table 1. Suppose the three cells with counts of 1, 3 and 6 are subject to masking methods.

Griffin, Navarro and Flores-Baez (1989) and Greenberg (1990) discuss two approaches to masking of tabular target data in this setting. *Suppression* deletes target variable information for the sensitive cells. *Imputation* replaces target data from cases in sensitive cells by imputed data from cases that match on some, but not all, the key variables. (A third approach discussed by these authors, *controlled rounding*, is described in Section 4.3.)

Table 2 presents data from Table 1 further classified by the target variable income. Table 3A shows a possible outcome of suppression applied to Table 2. The method retains the row and column margins of the table. The income values in the last column inside the table are suppressed (primary suppressions), and values in the first column are also suppressed so that the last column cannot be computed by subtraction; these are called secondary sup-

pressions. Columns for secondary suppressions are chosen so that the amount of additional information suppressed is minimized, although the details of implementing this rule become more complex in more elaborate cases (Greenberg 1990). Table 3B shows one possible outcome of imputation. Imputed incomes (marked with an asterisk) are taken from other records from different locations that match on S and R.

Suppression and imputation both delete target values (here income). The main distinction is that imputation fills in the deleted values, whereas suppression leaves the deleted values blank. As Griffin, Navarro and Flores-Baez (1989) point out, suppression shows more clearly that masking has taken place, but makes secondary analyses difficult. Imputed data are more readily amenable to aggregation and analysis by standard statistical software, and imputation is currently the option favored for 1990 census operations.

From an analysis perspective, note that masking by deletion is applied selectively here, so the masking selection mechanism is important. Since the decision to mask is based on sample values of the key variables (specifically, cell frequencies in the cross-classification by S, R and L), it is ignorable for suppression or imputation since values of key variables are not masked. However analyses (and imputation models) need to condition on the key variables to avoid bias. In particular, analyses of target variables (such as income) that effectively discard the suppressed cases are subject to bias. Also, imputation should be based on models (implicit or explicit) that properly reflect relationships between the key and target variables. For example, the imputation procedure in our example assumes that income is independent of location, conditional on race and household size. If this assumption is inadequate, then analyses of income are subject to bias, and the imputation method should be altered to reflect the relationship between the key and target variables more accurately.

Aside from the appropriate choice of imputation model, the primary analysis problem associated with imputed data is that the effective sample size is overstated since imputation error is not reflected in analysis of the filled-in data. Thus standard errors based on imputed data are understated, and tests and confidence intervals understate variability. Multiple imputation solves this problem while retaining much of the simplicity of analysis of imputed files. For more discussion of this point see Rubin and Schenker (1986); Rubin (1987); or Little (1988).

Suppression also differs from imputation in that it retains the original margins, and adds secondary suppressions to mask the target variable. The desire to retain the margins is understandable, and has the advantage that in principle it allows efficient analyses of models for which these margins are the sufficient statistics. However, secondary suppressions complicate matters considerably – Griffin, Navarro and Flores-Baez (1989) cite the problems of constructing secondary suppressions for a variety of data products. Imputation distorts the margins, but since the contribution from imputed cells is small the distortion should be minor.

The suppression and imputation methods discussed here mask the target variables. An alternative approach, which as noted in Section 1.2 may involve less information loss, is to mask the key variables, namely S, R and L in the example. Bethlehem et al. (1990) discuss this approach for microdata files that have uniqueness in key variables. Masking strategies such as collapsing or removing key variables are

Table 3. Table 2 masked by (A) suppression; (B) imputation; and (C) controlled rounding.

A. Suppressed data

		Size of household				
		1	2–3	4–6	>6	all
Income	< 30K	*S*	48	7	*P*	75
	30–60K	*S*	73	50	*P*	143
	> 60K	*S*	40	11	*P*	61
	all	44	161	68	6	279

Location = 1, Race = 2, *P* = Primary suppression, *S* = Secondary suppression.

B. Imputed data

		Size of household				
		1	2–3	4–6	>6	all
Income	< 30K	17	48	7	2*	74
	30–60K	18	73	50	4*	145
	> 60K	9	40	11	0*	60
	all	44	161	68	6	279

Location = 1, Race = 2, * = Imputed data.

C. Controlled rounding data, rounded to base of 5

		Size of household				
		1	2–3	4–6	>6	all
Income	< 30K	15	50	10	0	75
	30–60K	20	70	50	5	145
	> 60K	10	40	10	0	60
	all	45	160	70	5	280

Location = 1, Race = 2.

discussed, and may be reasonable if the key variables are not the major focus of analysis. Limiting the suppression of key variables to selected records can reduce the loss of information, but note that if selection is based on low frequencies in cross-tabulations of the key variables, the masking mechanism is non-ignorable. Resulting analyses that omit records with suppressed key variables are valid if the masked key variables appear as covariates, but are biased if the masked key variables appear as outcomes (Glynn and Laird 1986; Little 1992).

4.3. Masking by coarsening variables

4.3.1. Grouping or rounding categorical data

Grouping or rounding can increase confidentiality while retaining partial information on a response value. Two quite different applications of the idea appear in the masking literature: rounding counts in contingency tables, and rounding or grouping values of variables in a microdata file. The difference is illustrated in the next two examples.

Example 7. Controlled Rounding of Counts in a Contingency Table (Example 6 continued)

Controlled rounding rounds counts in crosstabulations to a suitable base (for example, 5), while preserving summation of internal counts to marginal totals. Table 3C illustrates the effect of controlled rounding with a base of 5 on the data in Table 2. Simple rounding of values in the table and the margins leads to inconsistency between the sums of the rounded values and the rounded margins. *Controlled* rounding modifies the rounded values inside the table to remove this inconsistency. The mathematical details are tricky; see for example Greenberg (1990). The effect of rounding on statistical analyses of contingency tables, such as loglinear models, appears to have received little attention. If the degree of rounding is minor, then standard analyses of the rounded counts may be adequate. If not, then the Poisson or multinomial error structure might be modified to

account for the addition of a component of error to the counts. Controlled rounding avoids the cumulation of rounding error in the margins, and hence improves unadjusted analyses based on these margins. It introduces a complicated correlation structure to the rounded counts that might have non-negligible effects on statistical inferences.

Example 8. Grouping Categorical Keys (Example 6 continued): Use of Fractional Records

Bethlehem et al. (1990) note that combining categories of key variables in Table 1 can reduce the incidence of uniqueness in a file, with some loss of information for analysis. The following extension limits the information loss for microdata files, and might be worthy of study. Cells with small counts are combined with adjacent cells to create a set S of C cells, labelled $\{1, 2, ..., C\}$ for convenience. Let m be the number of records in S, and write $m = \Sigma_c m_c$, where m_c is the number of records that originated in cell c, $c = 1, ..., C$. Then each case in S is replaced by C fractional cases, with fractional case c being assigned weight m_c/m and values of the key variables for cell c; values of target values and key variables not involved in S are unaltered. For example, consider the data in Table 1, and suppose the cells $(R, S, L) = (3, 4, 1)$, $(3, 3, 1)$ and $(2, 4, 1)$ are combined to form a set S with $1 + 3 + 6 = 10$ records. Each record in S is replaced by 3 records with weights 1/10, 3/10 and 6/10, $(R, S, L) = (3, 4, 1)$, $(3, 3, 1)$ and $(2, 4, 1)$ respectively, and the observed value of I and other target variables.

This microfile solves the uniqueness problem by breaking the link between the key and the target variables in sensitive cells. However (a) the full crosstabulation of the key variables can be constructed from the weighted data; (b) full information is retained for the target variables; (c) analysis involving both target variables and key variables not involved in the pooling into S is straightforward; (d) correct analysis involving both target variables and variables in S requires special missing-data methods.

4.3.2. Grouping quantitative outcomes

Grouping of a continuous variable can be useful at the data collection stage. For some sensitive quantitative variables (for example, income), a better response rate may be achieved by grouping the variable into categories (e.g., < 20K, 20K–40K, 40K–60K, 60K–80K, > 80K), and having the respondent provide the category rather than giving the exact value. Grouping can also be useful as a masking device to reduce disclosure risk. Direct analyses involving the grouped form of the variable follow standard lines for an ordered categorical variable. For analyses involving the underlying exact values of the variable, the literature on grouped or interval-censored data can be applied (e.g., Kulldorff 1961; Hasselblad, Stead and Galke 1980; Heitjan 1989).

Example 9. Masking by Grouping a Continuous Variable

Suppose $\mathbf{X} = (x_1, ..., x_n)^T$ is a random sample from a continuous distribution with density $f_X(x_i|\boldsymbol{\theta})$; for example, the normal distribution with mean θ_1 and variance θ_2. Let z_i represent a grouped version of x_i with J categories, defined by known cutpoints $\{c_j; j = 0, ..., J\}$, where $c_0 = -\infty$ and $c_J = \infty$; thus z_i given x_i has the degenerate distribution

$$f_Z(z_i = j|x_i) = \begin{cases} 1, & \text{if } c_{j-1} < x_i < c_j \\ 0, & \text{otherwise } (j = 1, ..., J). \end{cases}$$

The masking mechanism masks cases with

$z_i = j$ with known probability π_j

$$f_M(m_i = 1 | z_i = j, x_i) = \pi_j, \quad j = 1, ..., J.$$

This mechanism is ignorable since masking depends on values z_i that are known for all cases (whether or not they are masked). The likelihood of $\boldsymbol{\theta}$ given the masked data is

$$L(\boldsymbol{\theta}) = \prod_{m_i=0} f_X(x_i|\boldsymbol{\theta}) \prod_{m_i=1} \int_{I(z_i)} f_X(x|\boldsymbol{\theta})\, dx \tag{15}$$

where the region of integration $I(z_i)$ is (c_{j-1}, c_j) if $z_i = j$. In particular, suppose that cases with large values of x_i are vulnerable to disclosure. Setting $J = 2$, $\pi_1 = 0$ and $\pi_2 = 1$ results in all cases with $x_i > c_1$ being masked, yielding censored data with known censoring point c_1.

Likelihoods such as (15) can be analyzed by treating the true underlying values as missing and applying a missing-data technique such as EM; three examples are given in Little and Rubin (1987, Sec 11.3). Multiple imputation might also be used here, providing multiple imputes of the true value within the chosen interval. A simple technique that might yield satisfactory approximate inferences is to create multiple imputations by drawing randomly from a distribution within each interval; if the intervals are small then a uniform draw within the interval may be adequate (Heitjan and Rubin 1990; Little 1991). However, attention needs to be paid to the extreme categories. In the case of income, the lowest category is bounded by zero, but the highest category is unbounded. Imputation for the latter requires some information about the right tail of the distribution, at a minimum, an upper bounding value for the interval. In particular, analysis of highly skewed variables such as income can be very sensitive to the assumptions made about the right tail of the distribution (Rubin 1983). Imputing a random draw within the interval can also create distortions when the imputed variable is used as a predictor in regression, particularly when the variables are highly correlated. Sheppard's corrections can improve estimation in such settings (Dempster and Rubin 1983).

4.4. *Masking by noise injection*

One approach to masking is to delete and impute the sensitive values, as illustrated in Example 6. I have argued elsewhere (e.g., Little 1988) that imputation for missing data should be based on a model (implicit or explicit) for the predictive distribution of the missing values given the observed data; that the imputes should be drawn from this predictive distribution; and that the method of analysis of imputed data should reflect imputation error, using multiple imputation of the missing values (Rubin 1978b, 1987), or some other method. I use the term *predictive imputation* to describe imputation methods based on this philosophy.

Imputation for masking differs from imputation for missing data in an important respect: the underlying true values of the data are known to the data producer. Information about the true values might be supplied to the user in a form that does not compromise confidentiality, for example, by providing marginals as in Table 3A or the mean and covariance matrix of continuous data, or grouping variables as discussed in Section 4.3. This section discusses another way of exploiting information in the true values, namely imputing by adding noise. The relationship between predictive imputation and noise injection is examined in the following example.

Example 10. Adding Noise to Normal Data

Kim (1986) describes the following method for masking p continuous variables. Suppose the data consist of a random sample $\{\mathbf{x_i} = (\mathbf{x_{i1}}, ..., \mathbf{x_{ip}}) : 1 \leq i \leq n$ where $\mathbf{x}_i$ has mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma} = \{\sigma_{jk}\}$. Let $\bar{\mathbf{x}}$ and $\mathbf{S}$ denote the sample mean and covariance matrix, respectively, and let $\mathbf{e}_i^*$ denote a random draw from the p-variate normal distribution with mean $\bar{\mathbf{x}}$, covariance matrix $c\mathbf{S}$. Kim proposes replacing $\mathbf{x}_i$ by the masked vector $\mathbf{z}_i$, with jth component

$$\mathbf{z}_i = a(\mathbf{x}_i + \mathbf{e}_i^*) + (1 - a)\bar{\mathbf{x}} \qquad (16)$$

where a is chosen so that $\mathbf{z}_i$ has covariance matrix $\boldsymbol{\Sigma}$. Ignoring small sample corrections, $a = 1/\sqrt{1 + c}$. Sullivan and Fuller (1989) propose a similar method for normal variables, and extend it to non-normal and categorical variables by preliminary transformations to normality.

The following slight reformulation of this method provides a link with predictive imputation. Let $\mathbf{e}_i = \mathbf{e}_i^*/\sqrt{c}$ be a normal deviate with mean $\mathbf{0}$ and covariance matrix $\mathbf{S}$. If $\mathbf{x}_i$ was missing but $\bar{\mathbf{x}}$ and $\mathbf{S}$ were known, one method of predictive imputation replaces $\mathbf{x}_i$ by

$$\hat{\mathbf{x}}_i = \bar{\mathbf{x}} + \mathbf{e}_i \qquad (17)$$

the sample mean plus a random normal deviate. Consider masked values of the form

$$\mathbf{z}_i = \bar{\mathbf{x}} + \cos\phi n(\mathbf{x}_i - \bar{\mathbf{x}}) + \sin\phi n\mathbf{e}_i \qquad (18)$$

for $0 \leq \phi \leq \pi/2$. This method has the following properties

A. The first two moments of $\mathbf{z}_i$ match those of $\mathbf{x}_i$, ignoring $O(1/n)$ terms. More specifically, standard moment calculations yield

$$E(\mathbf{z}_i) = \boldsymbol{\mu}; \quad \mathrm{Var}(\mathbf{z}_i) = \boldsymbol{\Sigma}\{1 + (1 - \cos\phi)/n\}$$
$$\mathrm{Cov}(\mathbf{z}_i, \mathbf{z}_j) = \sin^2\phi n(\boldsymbol{\Sigma}/n)(i \neq j). \qquad (19)$$

B. $\mathbf{z}_i = \mathbf{x}_i$ when $\phi = 0$ and $\mathbf{z}_i = \hat{\mathbf{x}}_i$ when $\phi = \pi/2$. Hence this form of masking perturbs the true value towards a predictive imputed value (17); the angle ϕ provides an obvious geometric interpretation for this shift. The extreme case $\phi = \pi/2$ corresponds to replacing the true data $\{\mathbf{x}_i\}$ by an equal number of simulated cases $\{\mathbf{z}_i\}$ with the same mean and covariance matrix as the sample, a special case of the method in Example 11 below.

C. If $\mathbf{x}_i$ is multivariate normal, then $\mathbf{z}_i$ defined by (18) is also multivariate normal. If $\mathbf{x}_i$ is not normal, then the convolution of $\mathbf{x}_i$ with normal noise produces masked values that are more normal than the original values. Other forms of predictive imputation avoid the normal assumption by drawing from empirical distributions of residuals, here the set of values $\{\mathbf{x}_i - \bar{\mathbf{x}}\}$. However the convolution of the true values with similarly-distributed residuals still tends to distort the distribution of the masked values towards normality. Thus the method seems peculiarly suited to normal data, and indeed the extensions developed by Sullivan and Fuller (1989, 1990) involve transformations to normality prior to masking. These extensions appear promising but in need of further empirical validation.

D. From (A), $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and functions of these parameters can be consistently estimated by treating the masked values $\mathbf{z}_i$ as if they were the true values $\mathbf{x}_i$. However this approach does *not* yield valid inferences for parameters, since it does not account for the added uncertainty from masking. In particular inference for $\boldsymbol{\mu}$ is based on the masked sample mean $\bar{\mathbf{z}}$, which from (16) has mean $\boldsymbol{\mu}$ and covariance matrix

$$\mathrm{var}(\bar{\mathbf{z}}) = \frac{\boldsymbol{\Sigma}}{n}\left(1 + \sin^2\phi - \frac{\cos\phi(1 - \cos\phi)}{n}\right).$$

By comparison, the unmasked estimate $\bar{\mathbf{y}}$ has mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}/n$.

Hence correct inferences about means requires inflation of the usual estimate of standard error by a factor

$$\sqrt{1 + \sin^2 \phi - \cos \phi(1 - \cos \phi)/n},$$

which ranges from 1 when $\phi = 0$ to $\sqrt{2}$ when $\phi = \pi/2$. Similar factors could be worked out for analyses based on other parameters, such as functions of $\boldsymbol{\Sigma}$. These results imply that for valid point estimation the analyst does not need to be told the value of ϕ used by the masker, but for valid inferences the analyst needs to know ϕ, or variance inflation factors for the effects of masking.

An alternative approach to propagating the masking error is to multiply impute draws of $\mathbf{x}_i$ from the predictive distribution of $\mathbf{x}_i$ given $\mathbf{z}_i$, which is normal providing the original data are normal. The resulting data set can be analyzed using the simple methods described in Rubin (1987), and outlined in the appendix.

E. The masking method (18) can be generalized by replacing $\bar{\mathbf{x}}$ by the predicted mean given a set of covariates $\mathbf{c}_i$, and $\mathbf{e}_i$ by a draw from the residual (error) distribution of $\mathbf{x}_i$ given $\mathbf{c}_i$. In particular, categorical variables, which appear to fit rather awkwardly in this masking scheme, might be left unmasked or masked by some other method such as grouping, and noise-injection applied to the continuous variables conditional on the categorical variables, with predictive imputations based on a MANOVA model. Mixed strategies of this kind appear a fruitful topic for further research.

F. Point (D) shows that there is a penalty in information loss from noise injection. One might limit the loss by confining noise injection to a subset of cases that are vulnerable to disclosure. As noted in Section 3, a mechanism that selects based on the value of $\mathbf{x}_{i1}$, for example, censoring

$$f_M(m_i = 1|\mathbf{z}_i, \mathbf{x}_i) = \begin{cases} 1, & \text{if } x_{i1} > c, \\ 0, & \text{if } x_{i1} < c \end{cases}$$

is not ignorable, since x_{i1} is not observed for masked cases. The likelihood ignoring the masking mechanism is

$$L_1(\boldsymbol{\theta}) = \prod_{m_i=0} f_X(\mathbf{x}_i|\boldsymbol{\theta}) \times \prod_{m_i=1} \int f_Z(\mathbf{z}_i|\mathbf{x}) f_X(\mathbf{x}|\boldsymbol{\theta})\, d\mathbf{x}$$

which differs from the correct likelihood

$$L(\boldsymbol{\theta}) = \prod_{m_i=0} f_X(\mathbf{x}_i|\boldsymbol{\theta}) \times \prod_{m_i=1} \int_{x_i > c} f_Z(\mathbf{z}_i|\mathbf{x}) f_X(\mathbf{x}|\boldsymbol{\theta})\, d\mathbf{x}$$

where the integral is restricted to the region of censoring of $\mathbf{x}$. One ignorable selection scheme is to stratify on one variable and apply different masking rates across strata.

4.5. *Simulation artificial records*

An alternative to deletion or masking is to simulate artificial records and add them to the file, with or without inclusion of the original records. The following example shows implications for inference in the simple context of Example 10.

Example 11. Simulating Continuous Data

Suppose the data $\{\mathbf{x}_i : 1 \le i \le n\}$ are as in Example 10, and an additional m records $\{\mathbf{x}_i : n + 1 \le i \le n + m\}$ are simulated to have the same mean and covariance structure, using the predictive imputation method (17). Suppose now r of the original cases and the m new cases are included in the file, and write $\lambda = r/n$, $\delta = m/n$. The estimate of the mean $\boldsymbol{\mu}$ is then

$$\hat{\mu}_{\text{sim}} = (r\bar{\mathbf{x}}_1 + m(\bar{\mathbf{x}} + \bar{\mathbf{e}}))/n$$
$$= \frac{\lambda(1+\delta)\bar{\mathbf{x}}_1 + (1-\lambda)\bar{\mathbf{x}}_2 + \bar{\mathbf{e}}}{n(\lambda+\delta)}$$

where $\bar{\mathbf{x}}_1$ is the mean of the r included true values, $\bar{\mathbf{x}}_2$ is the mean of the $n-r$ excluded values, and $\bar{\mathbf{e}}$ is the mean of the added noise for the m simulated cases. This estimator is unbiased for μ with variance

$$\frac{\Sigma}{n}\left\{1 + \frac{\delta + \lambda(1-\lambda)}{(\lambda+\delta)^2}\right\}$$

ignoring higher order terms. Hence the proportional increase in variance over the unmasked estimator $\bar{\mathbf{x}}$ is $\rho = (\delta + \lambda(1-\lambda))/(\lambda+\delta)^2$, which provides a correction factor for inference based on $\hat{\mu}_{\text{sim}}$. For example, replacing the n cases by n simulated cases yields $\delta = 1$, $\lambda = 0$ and $\rho = 1$, or a 100% increase in variance; adding δn simulated cases and retaining all the original cases yields $\lambda = 1$, and $\rho = \delta/(1+\delta)^2$, which is near zero when δ is small or large and has a maximum of 0.25 when $\delta = 1$. Multiple imputation of simulated data is proposed in Rubin (1993).

Note that the adjustment to the variance requires knowledge of the fraction of simulated records in the file. If this analysis is extended to subsets of the data, then the proportion of simulated cases in the subset is also needed.

4.5. *Masking by reporting aggregate summaries*

It is clear that if data masked by predictive imputation, noise injection or simulation are analyzed in the same way as the unmasked data, some distortions result. An alternative approach is simply to provide directly the sufficient statistics for particular analyses, such as the sample mean $\bar{\mathbf{x}}$ and covariance matrix $\mathbf{S}$ in Examples 10 and 11. Since no changes are made to the data, analysis can be carried out using standard methods. Other information, such as the one-way marginal distributions of the variables, might also be provided if necessary. A challenge for previously-described masking procedures is to provide practically useful additional information over this simple approach.

The sample mean and covariance matrix are the usual set of summary statistics mentioned in applications, but this choice is clearly not appropriate for categorical variables. However, aggregation of data can also be applied in non-normal settings. From a modeling perspective, $\bar{\mathbf{x}}$ and $\mathbf{S}$ are sufficient statistics for the multivariate normal distribution. Sufficient statistics under other exponential family distributions are also suitable summaries, for example, low order marginal counts for a contingency table, which are sufficient for certain log-linear models, or the sample size, mean and covariance of a set of continuous variables within cells defined by a set of categorical variables. The latter allow fitting of regression models for continuous variables that include main effects of the crossclassifying variables, and interactions with those variables. Microaggregation, the presentation of aggregate information for sensitive cases, can be viewed as an extension of this idea, and seems worthy of more study.

Disadvantages of the approach include lack of flexibility in the choice of variables to be analyzed, and the relative inability to do exploratory analysis and model-checking. Also as a practical matter, analysis programs are needed that will accept the data in aggregate form. For example, current regression packages often allow input of the mean and covariance matrix, but log-linear model packages require modification to accept data in the form of low order marginals.

5. Concluding Remarks

There is clearly a close relationship between masked data and missing data, and hence the history of research in missing data provides an interesting perspective on the masked data problem. My own view (e.g., Little and Rubin 1987) is that missing-data research progressed from studies of ad-hoc "fixes" (such as fill-in methods) to more rigorous methods based on models for the data and missing-data mechanism. As the importance of masking for data confidentiality increases, there may be a parallel transition from ad-hoc approximate analysis, such as imputing confidential data and then treating it as the truth, to more careful analyses that take into account the masking process; the work of Kim (1986) and Fuller and his colleagues (Fuller 1993) provides important steps in this direction.

This paper has presented a model-based perspective for the analysis of masked data. The likelihood-based approach shows explicitly how masking activities can be modeled and incorporated into the analysis, and provides a rigorous basis for creating and assessing masking methods. Although a full likelihood-based analysis may not be feasible in many settings, I think the modeling perspective provides a useful basis for assessing simpler approximate methods. Future work might provide more detailed applications of the modeling approach to specific masking procedures.

Appendix

The EM Algorithm and Multiple Imputation

1. The EM Algorithm

As in Section 2.1, let $\mathbf{X} = \{x_{ij}\}$ denote an $(n \times p)$ data matrix prior to masking, $\mathbf{M} = \{m_{ij}\}$ denote the masking indicator matrix, with $m_{ij} = 1$ if x_{ij} is masked and $m_{ij} = 0$ otherwise. Let $\mathbf{Z} = \{z_{ij}\}$, where z_{ij} represents the masked value of x_{ij}, and write $\mathbf{X} = (\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{mis}})$, $\mathbf{Z} = (\mathbf{Z}_{\text{obs}}, \mathbf{Z}_{\text{mis}})$, where obs denotes observed components and mis missing components of each matrix. Given a model that specifies the distribution of $\mathbf{X}$ with density $f_X(\mathbf{X}|\boldsymbol{\theta})$, analysis of the masked data should be based on the likelihood for $\boldsymbol{\theta}$ given the data $\mathbf{M}$, $\mathbf{X}_{\text{obs}}$ and $\mathbf{Z}_{\text{obs}}$, which can be written in the form

$$L(\boldsymbol{\theta}|\mathbf{M}, \mathbf{X}_{\text{obs}}, \mathbf{Z}_{\text{obs}}) = \int f_X(\mathbf{X}|\boldsymbol{\theta}) \times f_Z^*(\mathbf{Z}_{\text{obs}}|\mathbf{X}) f_M(\mathbf{M}|\mathbf{X}, \mathbf{Z}_{\text{obs}})\, d\mathbf{X}_{\text{mis}}. \tag{A1}$$

Explicit ML estimates of $\boldsymbol{\theta}$ that maximize (A1) are often not available. Standard iterative algorithms such as Newton–Raphson or scoring can be used in such cases. However, the ubiquitous EM algorithm (Dempster, Laird, and Rubin 1977; Little and Rubin 1987, Ch. 7 and 11) is an alternative approach that can be easier to program, and provides insights into simpler incomplete-data methods based on imputation. Let

$$l(\boldsymbol{\theta}|\mathbf{M}, \mathbf{Z}_{\text{obs}}, \mathbf{X}) = \log\{f_X(\mathbf{X}|\boldsymbol{\theta}) f_Z^*(\mathbf{Z}_{\text{obs}}|\mathbf{X}) \times f_M(\mathbf{M}|\mathbf{X}, \mathbf{Z}_{\text{obs}})\}$$

denote the loglikelihood of $\boldsymbol{\theta}$ based on $\mathbf{M}$, $\mathbf{Z}_{\text{obs}}$ and hypothetical complete data $\mathbf{X} = (\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{mis}})$. Let $\boldsymbol{\theta}^{(t)}$ denote an estimate of $\boldsymbol{\theta}$ at iteration t of the algorithm. Iteration $t+1$ consists of an E-step and an M-step. The E-step consists of taking the expectation of $l(\boldsymbol{\theta}|\mathbf{M}, \mathbf{Z}_{\text{obs}}, \mathbf{X})$ over the conditional distribution of $\mathbf{X}_{\text{mis}}$ given $\mathbf{X}_{\text{obs}}$, evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$. That is, the expected loglikelihood

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \int l(\boldsymbol{\theta}|\mathbf{M}, \mathbf{Z}_{\text{obs}}, \mathbf{X}) \times f(\mathbf{X}_{\text{mis}}|\mathbf{X}_{\text{obs}}, \mathbf{M}, \mathbf{Z}_{\text{obs}}, \boldsymbol{\theta} = \boldsymbol{\theta}^{(t)})\, d\mathbf{X}_{\text{mis}}$$

is formed.

The M-step determines $\theta^{(t+1)}$ by maximizing this expected loglikelihood

$$Q(\theta^{(t+1)}|\theta^{(t)}) \geq Q(\theta|\theta^{(t)}), \quad \text{for all } \theta.$$

The new estimate $\theta^{(t+1)}$ then replaces $\theta^{(t)}$ in the next iteration. It can easily be shown that each step of EM increases the likelihood of θ given $\mathbf{X}_{\text{obs}}$. Also, under quite general conditions, EM converges to the maximum of this function. In particular, if a unique finite ML estimate of θ exists, EM will find it. If the masking mechanism is ignorable, then the distribution of $\mathbf{M}$ can be omitted; specifically, the E-Step computes

$$Q(\theta|\theta^{(t)}) = \int l(\theta|\mathbf{Z}_{\text{obs}}, \mathbf{X}) f(\mathbf{X}_{\text{mis}}|\mathbf{X}_{\text{obs}}, \mathbf{Z}_{\text{obs}}, \theta = \theta^{(t)}) d\mathbf{X}_{\text{mis}}$$

where

$$l(\theta|\mathbf{Z}_{\text{obs}}, \mathbf{X}) = \log\{f_X(\mathbf{X}|\theta) f_Z^*(\mathbf{Z}_{\text{obs}}|\mathbf{X})\}.$$

EM is particularly useful when the M-step is noniterative, or available using existing software. Note that the algorithm does not involve computing and inverting an information matrix at each iteration. This feature can be useful in problems with many parameters, since the information matrix is square with dimension equal to the number of parameters. Standard errors based on the inverted information matrix, however, are not an output of EM and hence if required need a separate computation. (Other methods of computing standard errors, such as profile likelihood or sample re-use methods, do not rely on the information matrix and may be preferable with moderate-sized samples.) Although EM is reliable in that it increases the likelihood at each iteration, it can be painfully slow to converge in problems where the fraction of missing information (defined in terms of eigenvalues of the information matrix) is large.

2. Multiple Imputation

Imputation creates a rectangular data set convenient for subsequent analysis by replacing masked values $\mathbf{X}_{\text{mis}}$ by estimates based on the masked data $(\mathbf{M}, \mathbf{X}_{\text{obs}}, \mathbf{Z}_{\text{obs}})$. Methods that supply a single impute are usually deficient in that they do not reflect imputation error; standard errors from the filled-in data are too optimistic. Rubin (1978, 1987) proposes multiple imputation as a solution for this problem. Two or more (say, m) values are drawn from the predictive distribution of the missing values, and then complete-data analyses are repeated m times, once with each imputation substituted. Let $\hat{\theta}_l$ be the estimate of a particular parameter θ from the lth analysis, and let $\hat{v}_l$ be the estimated variance. The final estimate of θ is $\hat{\theta} = \Sigma_l \hat{\theta}_l/m$, with estimated variance

$$\hat{v}^2 = s_w^2 + (1 + m^{-1}) s_b^2 \qquad \text{(A2)}$$

where $s_w^2 = \Sigma_l \hat{v}_l/m$ is the average variance within imputed data sets and $s_b^2 = \Sigma_l(\hat{\theta}_l - \hat{\theta})^2/(m-1)$ is the between-imputation variance, and reflects uncertainty in the imputation process. Large-sample inference for θ is based on treating $(\hat{\theta} - \theta)/\hat{v}$ as t distributed with $\nu = (m-1)[1 + \{m/(m+1)\}s_w^2/s_b^2]$ degrees of freedom. For theory underlying the method and practical examples, see Rubin and Schenker (1986) and Rubin (1987).

6. References

Bethlehem, J.G., Keller, W.J., and Pannekoek, J. (1990). Disclosure Control of Microdata. Journal of the American Statistical Association, 85, 38–45.

Bourke, P.D. and Moran, M.A. (1988). Estimating Proportions from Randomized Response Data Using the EM Algor-

ithm. Journal of the American Statistical Association, 83, 964–968.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum Likelihood from Incomplete Data Via the EM algorithm. Journal of the Royal Statistical Society, Ser. B, 40, 1–38.

Dempster, A.P. and Rubin, D.B. (1983). Rounding Error in Regression: The Appropriateness of Sheppard's Corrections. Journal of the Royal Statistical Society, Ser. B, 46, 51–59.

Duncan, G.T. and Lambert, D. (1986). Disclosure-Limited Data Dissemination. Journal of the American Statistical Association, 81, 10–28.

Duncan, G.T. and Lambert, D. (1989). The Risk of Disclosure for Micro-Data. Journal of Business and Economic Statistics, 7, 207–217.

Fuller, W.A. (1993). Use of Masking Procedures for Disclosure Limitation. Journal of Official Statistics, 9, 383–406.

Glynn, R.J. and Laird, N.M. (1986). Regression Estimates and Missing Data: Complete-Case Analysis. Technical Report, Department of Biostatistics, Harvard School of Public Health.

Greenberg, B. (1990). Disclosure Avoidance Research at the Census Bureau. Proceedings of the Annual Research Conference, Bureau of the Census, 144–166.

Griffin, R.A., Navarro, A., and Flores-Baez, L. (1989). Disclosure Avoidance for the 1990 Census. Proceedings of the Survey Research Methods Section, American Statistical Association, 516–521.

Hasselblad, V., Stead, A.G., and Galke, W. (1980). Analysis of Coarsely Grouped Data from the Lognormal Distribution. Journal of the American Statistical Association, 75, 771–778.

Heitjan, D.F. (1989). Inference from Grouped Continuous Data; A Review (with discussion). Statistical Science, 4, 164–183.

Heitjan, D.F. and Little, R.J.A. (1991), Multiple Imputation for the Fatal Accident Reporting System. Applied Statistics, 40, 13–29.

Heitjan, D.J. and Rubin, D.B. (1990). Inferences from Coarse Data Via Multiple Imputation; Age Heaping in a Third World Nutritional Study. Journal of the American Statistical Association, 85, 304–314.

Heitjan, D.F. and Rubin, D.B. (1991). Ignorability and Coarse Data. Annals of Statistics, 19, 2244–2253.

Ibrahim, J.G. (1990). Incomplete Data in Generalized Linear Models. Journal of the American Statistical Association, 85, 756–769.

Kim, J. (1986). A Method for Limiting Disclosure of Microdata Based on Random Noise and Transformation. Proceedings of the Survey Research Methods Section, American Statistical Association, 370–374.

Kulldorff, G. (1961). Contributions to the Theory of Estimation from Grouped and Partially Grouped Samples. Stockholm: Almqvist and Wiksell.

Little, R.J.A. (1988). Missing-Data Adjustments in Large Surveys. Journal of Business and Economic Statistics, 6, 287–301 (with discussion).

Little, R.J.A. (1991). Incomplete Data in Event History Analysis. In Demographic Applications of Event History Analysis, J. Trussell, R. Harkingson and J. Tilton, eds., Oxford: Clarendon Press, 209–230.

Little, R.J.A. (1992). Regression with Missing X's: A Review. Journal of the American Statistical Association, 87, 1227–1237.

Little, R.J.A. and Rubin, D.B. (1987). Statistical Analysis with Missing Data. New York: John Wiley.

Little, R.J.A. and Rubin, D.B. (1989). Missing Data in Social Science Data Sets. Sociological Methods and Research, 18, 292–326.

McGuckin, R.H. and Nguyen, S.V. (1990). Public Use Microdata: Disclosure and Usefulness. Journal of Economic and Social Measurement, 16, 19–39.

Paass, G. (1988). Disclosure Risk and Disclosure Avoidance for Microdata. Journal of Business and Economic Statistics, 6, 487–500.

Rubin, D.B. (1976). Inference and Missing Data. Biometrika, 63, 581–592.

Rubin, D.B. (1978a). Bayesian Inference for Causal Effects: The Role of Randomization. The Annals of Statistics, 6, 34–58.

Rubin, D.B. (1978b). Multiple Imputation in Sample Surveys – A Phenomenological Bayesian Approach to Nonresponse. Proceedings of the Survey Research Methods Section, American Statistical Association, 20–34.

Rubin, D.B. (1983). A Case Study of the Robustness of Bayesian Methods of Inference: Estimating the Total of a Finite Population Using Transformations to Normality. In Scientific Inference, Data Analysis and Robustness, New York: Academic Press, 213–244.

Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: John Wiley.

Rubin, D.B. (1993). Statistical Disclosure Limitation. Journal of Official Statistics, 9, (this issue).

Rubin, D.B. and Schenker, N. (1986). Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse. Journal of the American Statistical Association, 81, 366–374.

Sheppard, W.F. (1898). On the Calculation of the Most Probable Values of Frequency Constants for Data Arranged According to Equidistant Division of Scale. Proceedings of the London Mathematical Society, 29, 353–380.

Sullivan, G. and Fuller, W.A. (1989). The Use of Measurement Error to Avoid Disclosure. Proceedings of the Survey Research Methods Section, American Statistical Association, 802–807.

Sullivan, G. and Fuller, W.A. (1990). Construction of Masking Error for Categorical Variables. Proceedings of the Survey Research Methods Section, American Statistical Association, 435–439.

Warner, S.L. (1965). Randomized Response; a Survey Technique for Eliminating Evasive Answer Bias. Journal of the American Statistical Association, 60, 63–69.

Received September 1992
Revised April 1993