

# STATISTICAL NOTIONS OF DATA DISCLOSURE AVOIDANCE AND THEIR RELATIONSHIP TO TRADITIONAL STATISTICAL METHODOLOGY: DATA SWAPPING AND LOGLINEAR MODELS

Stephen E. Fienberg and Russell J. Steele, Carnegie Mellon University  
Udi Makov, Haifa University

## ABSTRACT

For most data releases especially those from censuses, the U. S. Bureau of the Census has either released data at high levels of aggregation or applied a data disclosure avoidance procedure such as data swapping or cell suppression before preparing micro-data or tables for release. In this paper, we present a general statistical characterization of the goal of a statistical agency in releasing confidential data subject to the application of disclosure avoidance procedures. We use this characterization to provide a framework for the study of data disclosure avoidance procedures for categorical variables.

Consider a sample of  $n$  observations on  $p$  variables, which may be discrete or continuous. Our general characterization is in terms of the smoothing of a multi-dimensional empirical distribution function (an ordered version of the data), and sampling from it using bootstrap-like selection. Both the smoothing and the sampling introduce alterations to the data and thus a bootstrap sample will not necessarily be the same as the original sample -- this works to preserve the confidentiality of individuals providing the original data. Two obvious questions are: How well confidentiality is preserved by such a process? Have the smoothing and sampling disguised fundamental relationships among the  $p$  variables of interest to others who will work only with the altered data? Rubin (1993) has provided a closely related characterization and approach based on multiple imputation.

We explain some of these ideas in greater detail in the context of categorical random variables and compare them to methods in current use for data disclosure avoidance such as data swapping and cell suppression. We also relate this approach the data disclosure avoidance methods to statistical analysis associated with the use of loglinear models for cross-classified categorical data.

## KEYWORDS

Bootstrap, Cell suppression, Confidentiality, Contingency table analysis, Data-swapping, Graphical models, Multiple imputation.

## 1. INTRODUCTION

Disclosure avoidance methodology has developed over the past 20 years as a major area of government statistics research and activity. The advances are impressive (e.g. see the progress chronicled in Subcommittee on Disclosure-Avoidance Techniques, 1994, especially when compared with the methodology described as of Subcommittee on Disclosure-Avoidance Techniques, 1978), but all too often they appear to be unlinked to the analytical uses to which most census and survey data are put and to the evolving methods of analysis. During this same 20 year period there have also

been major advances in statistical methodology and theory. A theme of this paper is that many of these statistical tools that come from these latter developments have relevance to the area of disclosure avoidance methodology. For a number of reasons situations involving categorical data in the form of a contingency table offer an excellent venue for such consideration.

In this paper we:

- Review some current statistical ideas in use for data disclosure avoidance for categorical variables.
- Present a new statistical framework for data release.
- Relate these ideas and approaches to "traditional" statistical methodology associated with loglinear models for cross-classified categorical data.

Before doing so, we outline a framework in which the problem of data-disclosure avoidance methodology can be viewed. Consider four different parties:

- The *Agency* or data collector.
- The *Respondents* or data providers.
- An *Intruder* who wants to learn about one or more data providers via the data to be released by the agency.
- *Users* or secondary analysts of the agency data.

The question of interest to us is: What data can the agency release for analysis by the users while protecting the respondents from the intruder (i.e., preserving their confidentiality)? The practical way in which this question has been answered is through the application of some disclosure limitation method that the agency hopes achieves the desired goals.

In the next section we review some of the specific methods for disclosure avoidance that have been proposed in the literature, and that fit under the broad rubric of "matrix masking." In particular we describe two specific methods for "matrix masking" when all of the variables are categorical -- cell suppression and data swapping. Then, in Section 3, we explain how we view these methods in the context of the users' analytical goals. In Section 4, we suggest a general strategy for disclosure limitation that attends to the proposed goals in a non-standard fashion, and we relate the strategy to some modern approaches from the statistical methodology literature. In Section 5, we describe in further detail how we propose implementing the strategy in the context of contingency table problems. We end by outlining research that would put the general strategy suggested on a firm theoretical foundation.

There are a number of excellent papers that attempt to bridge the gap between the literature on disclosure avoidance and more general statistical methodology, beginning with the pioneering work of Duncan and Lambert (1986, 1989), and continuing with Fuller (1993), Lambert (1993), Rubin (1993) and other contributors to a special issue of the *Journal of Official Statistics*. This paper builds, both directly and indirectly on a number of these earlier efforts.

The general strategy proposed here has appeared in other papers in the past, e.g. see Liew, et. al. (1985) and Rubin (1993), and Fienberg (1994b), and Heer (1993) has suggested a bootstrap method for contingency tables which is related but different from our proposals in Section 5. To our knowledge, no previous authors have integrated these ideas with both the full literature on loglinear

model methods and that on disclosure limitation.

## 2. MATRIX MASKING FOR MICRO-DATA

Duncan and Pearson (1991) give an excellent description of approaches to the masking of microdata. Suppose that  $X$  is an  $n$  by  $p$  matrix representing the microdata for  $n$  individuals or cases on  $p$  variables or attributes. Then matrix masking of the microdata file  $X$  provides the user with the transformed file  $Z=AXB+C$  in lieu of  $X$ . The matrix  $A$  transforms cases,  $B$  transforms variables, and  $C$  blurs the entries of  $AXB$ . The use of  $Z$  in lieu of  $X$  includes several well-known approaches as special cases:

1. Release a subset or sample of the data (delete rows of  $X$ ).
2. Include simulated data (add rows to  $X$ ).
3. Add random perturbations to  $X$ .
4. Exclude selected attributes (delete columns of  $X$ ).
5. Release the variance-covariance matrix (choose  $A = X^T$ ).

Examples of transformations to  $X$  that are not of the form  $Z$  include swapping (exchanging rows for a subset of the columns of  $X$ ) and the coarsening, grouping or truncation of attributes. But Cox (1994) explicitly denotes these methods, especially swapping (see below and the Appendix) to the matrix masking approach.

Clearly the use of  $Z$  needs some information about  $(A, B, C)$ , but the release of full information is not allowed. Determining what information can be released for a given choice of  $(A, B, C)$  and the choice of  $Z$  itself are both active areas of research. For further details see Cox (1994) and Fienberg (1994a) as well as the specific work of Fuller (1993) and Sullivan (1989).

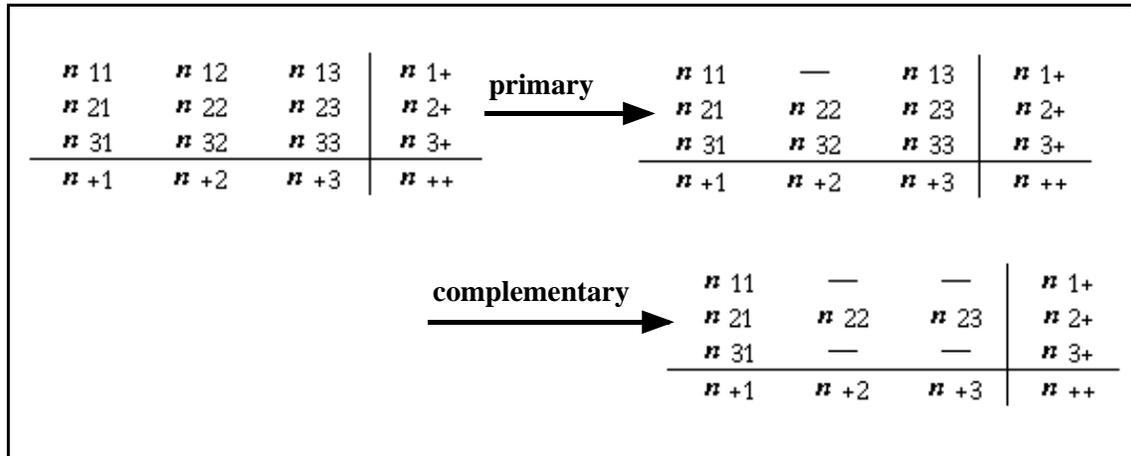
A special case involving the deletion of rows is the method of cell suppression. Suppose we are interested in summarizing a set of data in the form of a cross-classification of counts or nonnegative aggregates. Deleting or suppressing a cell value is equivalent to the deletion of those rows of  $X$  for which the entries in columns corresponding to the cross-classifying variables assume the values that specify the cell in question. Cell suppression is widely used for data on establishments because counts of "1" or "2" may uniquely identify a respondent.

Current practice at the U. S. Census Bureau is to suppress any cell where  $k \geq 3$  or fewer respondents make up that cell's value. Such cells are referred to as *primary* suppressions. The bureau keeps the value of  $k$  as well as the method used for selection of cells confidential.

Because reported cross-classifications usually include the corresponding marginal totals, suppressing a single cell produces multiple masks for the same matrix and, taken together, these masks do not disguise the data -- the value of a deleted cell in a two-way array can be retrieved from the other entries in the same row or column combined with the corresponding marginal total. Thus methods for cell suppression in cross-classifications also choose other cell values for suppression; these are often referred to as *complementary* suppressions. Determining "desirable" patterns of complementary suppressions is an active area of research, especially for multi-way cross-classifications (e.g., see Greenberg and Zayatz, 1992; Cox, 1980, 1995; and Carvalho, et al., 1994, Robertson, 1993).

In Figure 1 we depict an example of cell suppression in a two-way contingency table with entries

$\{n_{ij}\}$  involving a single primary suppression, in the (1,2) cell, and three complementary suppressions. It is important to note for the present context that the basic approach is one involving margin preservation, i.e., in the two-way table the method for suppression preserves both sets of one-dimensional marginal totals,  $\{n_{i+}\}$  and  $\{n_{+j}\}$ , by design. In higher dimensions cell suppression also preserves marginal totals but possibly those of highest order.



**Figure 1:** An Illustration of Cell Suppression in a Two-way Contingency Table with Entries  $\{n_{ij}\}$ . The primary suppression occurs in cell (1,2) and the complementary suppressions in cells (1,3), (3,2), and (3,3).

In 1978, Dalenius and Reiss proposed a method for *swapping* observations while preserving marginal totals. In the Appendix we provide some details on their proposal and two variations on the same theme, including the one used for disclosure limitation in micro-data files from the 1990 U. S. decennial census. Again this we can view data swapping as a special case of matrix masking at least in its simplest forms as noted above.

In Figure 2 we present an illustration of data swapping in a three-way contingency table, in which an observation for the (1,2,1) cell is moved to the other layer, i.e. into the (1,2,2) cell, and in return an observation from the (3,1,2) cell is moving to the first layer, i.e., to the (3,1,1) cell. Thus in moving from the original table (on the left) to the table with the swapped pair of observations (on the right) we end up preserving the two-way totals,  $\{n_{ij+}\}$ , and the one-way totals,  $\{n_{+jk}\}$ . Data swapping involves the repeated application of such movements of pairs of observations.

$n_{111}$	$n_{121}$	$n_{1+1}$	$n_{111}$	$n_{121} - 1$	$n_{1+1} - 1$
$n_{211}$	$n_{221}$	$n_{2+1}$	$n_{211}$	$n_{221}$	$n_{2+1}$
$n_{311}$	$n_{321}$	$n_{3+1}$	$n_{311} + 1$	$n_{321}$	$n_{3+1} + 1$
$n_{+11}$	$n_{+21}$	$n_{++1}$	$n_{+11} + 1$	$n_{+21} - 1$	$n_{++1}$
<p style="text-align: center;"><b>swapping layers for (1,2,1) &amp; (3,1,2)</b></p> <p style="text-align: center;">→</p>					
$n_{112}$	$n_{122}$	$n_{1+2}$	$n_{112}$	$n_{122} + 1$	$n_{1+2} + 1$
$n_{212}$	$n_{222}$	$n_{2+2}$	$n_{212}$	$n_{222}$	$n_{2+2}$
$n_{312}$	$n_{322}$	$n_{3+2}$	$n_{312} - 1$	$n_{322}$	$n_{3+2} - 1$
$n_{+12}$	$n_{+22}$	$n_{++2}$	$n_{+12} - 1$	$n_{+22} + 1$	$n_{++2}$

**Figure 2:** An Illustration of Data Swapping in a Three-way Contingency Table with entries  $\{n_{ij}\}$ . The original table is on the left and the table with observations from the (1,2,1) and (3,1,2) cells, swapped between layers is on the right.

Both the method of cell suppression and the method of data swapping preserve marginal totals in contingency tables. But this is also a property associated with loglinear model methods. What is interesting is that despite the fact that cell suppression and data swapping have been presented in the same sessions in various forums (e.g. see Cox and Sande, 1978 and Dalenius and Reiss, 1978 and the discussion of the two papers by Zalkind, 1978), previous authors have failed to note this clear relationship between these methods as well as to methods in the contingency table literature.

### 3. PERSPECTIVE ON DATA RELEASE AND DISCLOSURE LIMITATIONS

#### 3.1 The Users' Perspective

Typical users of government statistical data are interested in relationships and causal connections for policy choices. They use statistical models to describe such relationships. Often their view of "error" is akin to including an error component in an analytical model (e.g., such as a regression error term  $\varepsilon$  in the equation  $Y = \beta_0 + \beta_1 X + \varepsilon$ ). Otherwise, the typical user has limited ways to address the multiplicity of information on uncertainty and error coming from the statistical agency that produces the data.

The typical user is interested in analytical models and especially ones with causal implications. Thus we can think of the users' objectives as involving the linking of response variables,

$Y$ , and explanatory variables,  $X$ , through a statistical model that attempts to represent some underlying substantive phenomenon. Unfortunately we rarely get to observe or measure  $Y$  and  $X$  directly. What is produced through a census or a survey questionnaire is often a related but fallible measure of the quantities of real interest. These we label  $Y^*$  and  $X^*$ .

The user is interested in models for the conditional distribution of  $Y$  given  $X$  and thus we can take as the user's objective the estimation of a multivariate cumulative distribution function (c.d.f.), of the forms  $F_{Y|X}$  or  $F_{Y|X,\theta}$  for various values of  $X$ , or at least characteristics of such a multivariate c.d.f. Here the parameter  $\theta$  might be a population mean or variance,  $\mu$  or  $\sigma^2$ , or a parameter(s) in a statistical model such as a regression coefficient,  $\beta$ , likely multidimensional in form. While there has been some interest in the survey literature in the problem of estimation distribution functions (e.g., see Rao, 1994, and the references contained therein), although this literature has been concerned primarily with univariate  $Y$ . In the ensuing discussion we ignore those sources of measurement error in  $X$  beyond those forms captured in the agency's own evaluation and data preparation activities.

Estimation of a multivariate c.d.f. is a general statistical problem that includes a number of interesting special cases. For example, suppose that all of the variables in the user's model and in the data set are categorical in nature, as is often the case in census and survey settings. Then the c.d.f. is essentially equivalent to the table of conditional probabilities (for  $Y$  given  $X$ ) that correspond to the cross-classification of the variables in contingency table form (c.f., Bishop, Fienberg, and Holland, 1975). We refer to this special case again in the Section 5 and provide an extended set of references and notes on this special case.

### 3.2 The Current Agency Approach

At the risk of oversimplification, we can characterize the standard approach to data collection, processing and release roughly as follows:

- Collect and "clean up" the raw data. This includes editing, matching and all other preliminary processing.
- Protect the data by applying some form of data disclosure avoidance methodology.
- Then release the resulting data in one or perhaps both of the following forms:
  - as set of marginal tables for some larger cross-classification (i.e., selected marginal cross-classification -- see the discussion in Sections 2 and 5 regarding the relationship between marginal tables and loglinear models).
  - as micro-data files for the variables related to the ones of user interest ( $Y^*, X^*$ ).
- Estimate  $\theta$  directly using a sample-based quantity,  $\bar{\theta}$ .

In effect, the user then follows the agency's lead and estimates the c.d.f., directly from the released data using the "empirical" c.d.f. (suitably weighted to take into account the impact of the survey design),  $\bar{F}_{Y^*|X^*}$ , or possibly a more elaborate and smoother parametric estimate based on the estimated parameter, i. e.  $\bar{F}_{Y^*|X^*, \bar{\theta}}$ .

### 3.3 Shortcomings of The Current Approach

While this approach might make considerable sense for some descriptive statistical problems, the fact that  $F_{Y|X}$  and  $F_{Y|X,\theta}$  rarely reflect fully aspects of sampling design error that many believe to be important, such as clustering, and they almost never reflect the other sources of error listed above that typically dwarf sampling error. Further, given the relatively primitive statistical state of disclosure avoidance methodology, the user may still be able to "identify" individuals in the released data. One way to overcome these shortcomings is to continue to address the various components of error and to separately improve the approach to data disclosure avoidance. Alternatively, we can attempt to reconceptualize the data reporting problem in a new and integrated fashion.

## 4. A NEW STRATEGY AND FRAMEWORK

In this section, we propose a new approach to the release of survey data. We begin with the goals of the users and ask how agencies should organize the data of interest in order to provide data released that fit with the user goals.

### 4.1. Generating "Pseudo" Micro-Data Files for Public Use

Our new approach is cast in terms of the release of a public-use micro-data file that is intended to support analyses for the conditional distribution of  $Y^*$  given  $X^*$ . The first step in our prescription is:

1. Combine the census or survey data that the agency would normally have chosen to release, in the form  $\bar{F}_{Y^*|X^*}$  and  $\bar{F}_{Y^*|X^*,\bar{\theta}}$ , with formal statistical information on error, e.g., form editing, matching, nonresponse, etc. and apply some form of parametric or semi-parametric technique to estimate  $F_{Y|X}$  and  $F_{Y|X,\theta}$  by  $\hat{F}_{Y|X}$  and  $\hat{F}_{Y|X,\hat{\theta}}$  respectively, where  $\hat{\theta}$  is a new estimate of  $\theta$  cast in terms of the distribution of the variables of actual user interest,  $Y$  and  $X$ .

For non-parametric estimation of  $F_{Y|X}$  we can either think in terms of a classical statistical approach using some type of kernel density estimator or a related type of "smooth" estimate (e.g., see Scott, 1992), or a Bayesian approach based on the mixture of Dirichlet processes (e.g., see West, Müller, and Escobar, 1994; Gelfand and Mukhopadhyay, 1995) or the use of Polya trees (Lavine, 1992). These tools, however, have been used primarily in low-dimensional problems and thus there needs to be additional research to study their adaptation to the high-dimensional census and survey problems which are the focus of this paper. Even if these methods are not especially efficient for statistical estimation purposes, they may serve the needs of data disclosure avoidance which are crucial to the strategy outlined here.

In what ways does this new smoothed estimate of  $F_{Y|X}$  differ from the one that is explicit or implicit in the current approach? We offer three examples. First, consider the release of census data. In both the US and Canada, there has been extensive documentation of the extent of census undercoverage and how the resulting undercount is distributed across groups in the population and across geographic areas. Failure to correct for such undercoverage in the release of data of the form leads to biased estimates of the true quantity of interest,  $F_{Y|X}$ . Second, by smoothing data to reflect regression-like relationships we can typically achieve improved estimates with much lower

variances, although at the price of some potential bias. Finally, by incorporating agency information on components of error (which tends to increase variances) into the statistical estimation process, we produce a new smoothed estimator of  $F_{Y|X}$ .

The next steps in our prescription are:

2. Instead of releasing the c.d.f. estimate in step 1 above, the agency now "samples" from it to create a "pseudo" micro-data file which we label as  $\widehat{F}_{Y|X}$  and  $\widehat{F}_{Y|X,\hat{\theta}}$ . We use the overbar to indicate a sample from the smoothed c.d.f.'s, in accord with our earlier notation for the empirical c.d.f., which corresponds to a sample and the hat to indicate that we are sampling from the smoothed or estimated c.d.f.).
3. The agency repeats the process of "sampling" and then releases the resulting replicate "pseudo" micro-data files.

#### 4.2 Features of Pseudo Micro-Data File

The "pseudo" micro-data files created in the approach outlined above have several interesting features. First, if we think of  $\widehat{F}_{Y|X}$  and  $\widehat{F}_{Y|X,\hat{\theta}}$  as consisting of a set of released records for individuals, then these "individuals" do not necessarily correspond to any of those individuals in the original sample survey. This enhances the public notion of the protection of confidentiality of responses even if an intruder might still be able to indirectly make inferences about individuals in the original sample.

This point is especially important from the perspective of data disclosure avoidance. Since the individuals in the pseudo micro-data file are not necessarily those from the original sample, we have at least in part addressed confidentiality concerns. After all, we no longer even appear to be releasing data for any individual from the original sample. But this discussion of data disclosure avoidance is somewhat illusory. It remains possible that individuals, whose values on  $Y$  and  $X$  are far from those for the rest of the sample, may still in effect be regenerated through this complex statistical estimation process and reemerge virtually intact in the pseudo micro-data file. Thus we would argue that empirical checks on the effectiveness of data disclosure avoidance are still necessary and, in particular, we would advocate examining the issue from the perspective of an intruder (e.g., see Fienberg, Makov, and Sanil, 1994).

Second, there is close connection here with two recently developed statistical methods: (1) the bootstrap (Efron, 1979, Efron and Tibshirani, 1993, Hall, 1992) which is a classical method involving repeated sampling (with replacement) from an empirical distribution function; (2) multiple imputation (Rubin, 1987, 1993) which is a Bayesian method for generating values that are sampled from a posterior distribution. Our preference is to think about the estimation implicit in the approach outlined here from a Bayesian point of view. Thus, in effect, we are proposing that agencies should first estimate the empirical distribution function, generating the full posterior distribution of  $F_{Y|X}$  or  $F_{Y|X,\theta}$  and then sample from it using Rubin's multiple imputation approach. From this perspective, the bootstrap can be viewed as a way to sample from something approximately akin to the mean of the posterior distribution.

Third, the sample design for the released records need not be the same as that for original sample survey. Thus, at least in principle, the agency could use simple random sample or even sampling with replacement from  $\hat{F}_{Y|X}$  or  $\hat{F}_{Y|X, \hat{\theta}}$ . Rubin (1993) emphasizes this point without explaining exactly how to determine what we might call the "equivalent" sample size for the released data files. The heuristic idea is that there is only so much information available in the data and the resampling process cannot increase this. To preserve the appropriate level of accuracy in the data we need to have a bootstrap sample size that at least is conceptually equivalent to the "effective sample size" of the complex sample design, thus reflecting a design effect. This notion is somewhat problematic, however, as the "effective sample size" might well vary from one analytical setting to another!

But perhaps the most important feature of the approach is that users can now analyze pseudo micro-data files to estimate specific quantities of interest, e.g.  $\theta$  using standard statistical methodology. In essence the idea is that we can use a standard statistical method such as regression analysis or something more elaborate and thus will produce consistent estimates of the coefficients of interest. What we cannot do, however, is use the usual estimates of standard errors that result from the standard analysis tools. One of the lessons from both the bootstrap and multiple imputation is that while we can estimate  $\theta$  using standard statistical methodology applied to the generated bootstrap or multiple imputation sample, we cannot get a proper handle on the variability of our estimates without using replicate versions of the pseudo micro-data file. Generating multiple replicates, however, is a relatively simple task and estimating variances using the multiple versions of estimated parameters is then straightforward and, doesn't necessarily require special computer programs.

## 5. SOME DETAILS FOR THE CATEGORICAL DATA CASE

Here we outline the estimation and simulation process of Section 4 for the special case of categorical variables and cross-classification. Our focus is on parametric estimation of the c.d.f. which as we note above is equivalent to estimating the cell probabilities in a contingency table.

The most common class of statistical models used in connection with contingency table data is the loglinear model and for a set of basic sampling schemes (e.g., see Bishop, Fienberg, and Holland, 1975 and Fienberg, 1980) there is a direct relationship between a specific hierarchical loglinear model and a set of marginal tables that correspond to the minimal sufficient statistics associated with the model. If we report only those marginal totals appropriate for a loglinear model that fits the data well, then another investigator can, in effect, reconstruct the cell probabilities for the full contingency table (c.f., Fienberg, 1975). Further, reporting only a specific set of marginal tables is saying that these are the only totals needed for inference and this is implicitly suggesting the appropriateness of a specific loglinear model.

As we noted in Section 2, the two most commonly used methods for data disclosure avoidance in categorical variable settings are cell suppression and data swapping. Unfortunately there seems to be a total disconnect between the literature on disclosure avoidance for categorical variables and the now standard literature on loglinear models for categorical data. This is rather unfortunate since, as we noted in Section 2, the notion of margin preservation is fundamental to both cell suppression and data swapping. In the former, cells are suppressed subject to marginal constraints and, in the latter, individuals with one set of margins fixed are swapped between cells thus preserving other totals. Thus key features of these methods can be embedded in the loglinear model framework thus suggesting alternative ways to approach disclosure avoidance. Further results from the loglinear model literature may well be of value in understanding the properties of methods such as cell

suppression and data swapping (c.f. the discussion in Fienberg, 1995), but here we pursue an alternative approach linked to the general strategy described in Section 4.

Finding a cross-classified table of counts that satisfies a given set of marginal constraints is a problem which has occupied the attention of a substantial number of statisticians in recent years (e.g., see Agresti 1993; Zelterman, Chan, and Mielke, 1995). A number of algorithms have been proposed but they have been implemented primarily for two- and three-way cross-classifications (e.g., see Patefield, 1981). New ideas from the literature on graphical loglinear models suggest that implementation for higher dimensions may at last become feasible (e.g., Edwards, 1995, and Lauritzen, 1996, or Whittaker, 1990 for details on graphical models). The framework we outline in Section 3 requires us to produce a smooth c.d.f. and then sample from it. In the present context, this seems to suggest, at least heuristically, that we should consider making draws from the exact distribution conditional on a fixed set of marginal totals.

Consider a three dimensional contingency table with cell counts  $\{n_{ij}\}$  and expected cell values  $\{m_{ijk}\}$ . We can fit loglinear models to the expected cell values such as the model of no 2nd-order interaction

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(jk)} + u_{23(jk)} \quad (1)$$

with appropriate side-constraints for identification purposes. The minimal sufficient statistics or “fully efficient statistics” for this model are the margins that correspond to highest order terms:  $\{n_{ij+}\}$ ,  $\{n_{+jk}\}$ ,  $\{n_{i+k}\}$ .

A special case of model (1), in which  $u_{23(jk)}=0$  for all  $j$  and  $k$  interpretable as the conditional independence of variables 2 and 3 given variable 1. All conditional independence models for a multidimensional contingency table are loglinear models.

Darroch, Lauritzen, and Speed (1980) introduced the special subfamily of loglinear models known as graphical loglinear models, which are characterized by simultaneous conditional independence relationships. This subfamily of models can be represented by a set of graphs whose nodes correspond to the variables of the table and where the absence of an arc connecting two nodes implies that those variables are conditionally independent given the remaining variables.

In unpublished work in the 1970’s, Darroch attempted to construct a Markov chain algorithm for generating draws from the conditional distribution given the margins implied by a loglinear model. His transitions in effect involved one-step data swaps. Glonek (1987) showed that the resulting algorithm converges only when the Markov chain is irreducible. In particular, he showed that this was not the case for the no 2nd-order interaction model, (1), for a 3-way table.

Diaconis and Sturmfels (1993) showed how to implement the Darroch-Glonek approach and provide a proof of the convergence of the algorithm through the irreducibility of the Markov chain. We propose to generate draws from the exact distribution under graphical loglinear models using their algorithm. In order to ensure some level of smoothness in the resulting tables, we can retain only those draws “compatible” with a more complex loglinear model..

Alternatively we can generate a full posterior distribution of the cell probabilities in the table, e.g., using the methods of Epstein and Fienberg (1992) or Madigan and York (1995), and then sample from that posterior distribution.

We are in the process of actually implementing this strategy using data from the 1990 decennial census and the Diaconis-Sturmfels-Glonek-Darroch algorithm for graphical models.

## **6. TAKING VARIABILITY SERIOUSLY**

It is important to distinguish between the idea of generating public-use micro-data files based on real people and real data through a statistical simulation process, such as we have outlined in this paper, and the typical micro-simulation model, which may rely on indirectly on data via statistical models but which does not correspond to data on real people. There is a serious difference between "pseudo people" who resemble individuals from whom we have actually collected data of interest, and "imaginary people" for whom we have invented data through a stochastic or nonstochastic modeling process. In this paper we propose the former, not the latter.

### **6.1 Virtues of Proposed Framework**

There are several virtues of the proposed framework outlined above. First, we believe that it would force agencies to take their own data and their sources of error more seriously, as these are key inputs to the modeling effort outlined in Section 4. Second, we believe that it would solve a large part of the data disclosure avoidance problem. Third, the framework would generate public-use micro-data files of a form that would allow users to apply standard statistical methodology and model search methods.

### **6.2 Examples of Research to Be Done**

There are a number of formidable technical details that need to be addressed before an agency could properly implement the proposed framework. Examples of these include:

- How should an agency combine the multiple sources of error and uncertainty?
- What smoothing methods should be used and how much smoothing is appropriate?
- How do we determine "effective" sample size for pseudo micro-data files? The application of bootstrap ideas relies on certain series expansions (e.g., see Hall, 1992), and these typically require the use of a bootstrap sample of the same size as the original sample. What is the equivalent notion here?
- How many replicates are required for variance estimation? Rubin (1987, 1993) suggests the use of four or five replicates in the multiple imputation context. Efron and Tibshirani (1993) uses large numbers of bootstrap replications. Will a smaller number suffice for either approach?

Further the actual implementation of algorithms of the highly multidimensional situations involved in censal and survey data may require new statistical methods and theory. For example, as we suggested in Section 5, the problem of simulating from distributions for multidimensional contingency tables subject to marginal constraints has been implemented primarily for two and three-dimensional tables. Implementation for higher dimensions requires new strategies and algorithms. These are at the forefront of current statistical and mathematical research.

Finally, we may need to think about the statistical estimation problems outlined here in a form different from that which we usually find in the methodological literature. Because of the multiplicity of goals that we are attempting to address, we may need to think in terms of providing the users with data that enable them to approximate the conditional distributions  $F_{Y|X}$  and  $F_{Y|X,\theta}$  and rather than reproduce them in a more precise statistical fashion. This relates to Meng's (1994) notion of uncongeniality between an imputer's assessment and those assessments of the users.

### **6.3 Summary**

In this paper, we have tried to suggest that both government agencies and users bear responsibility when it comes to utilizing census and survey data. It is no longer enough for agencies to prepare public-use files and extensive sets of tabulations as they have in the past. Nor can they continue to ignore the analytical goals of the users of their data. At the same time, the users must learn how various sources of survey error affect their analytical goals, and to build such information into the statistical procedures they use.

We have argued that, by looking to and utilizing recent developments in statistical methodology, we may be able to develop an integrated approach to the release and analysis of survey data which will help us all learn to take uncertainty and error seriously. Perhaps the framework proposed in this paper will be the first step towards this goal.

## **APPENDIX: AN OVERVIEW OF DATA SWAPPING AND ITS METHODS**

### **1. Data Swapping Methods**

The method of data swapping was originally proposed by Dalenius and Reiss (1978) and subsequently considered by several others. A variant of the method was used in the 1990 U.S. decennial census. In this Appendix we briefly describe this work, especially in terms of how it can be implemented with categorical data in the form of a multi-dimensional contingency table.

#### **Dalenius / Reiss Version**

Data-swapping, according to Dalenius and Reiss (1982), is a way of presenting usable database information, without compromising the security of any single piece of data. It involves swapping the locations of individual pieces of data in a database in such a way that certain underlying statistics remain unaltered.

The driving methodology behind data-swapping is contained at the beginning of Section 4 of Dalenius and Reiss (1982):

The basic idea is that the value of a sensitive variable for a particular individual cannot be compromised if there are at least two distinct databases that are consistent with the underlying statistics and that assign different values to that variable. This notion was extended to a complete database by noting that a database is protected if and only if each of the sensitive values is protected.

Therefore, if there exist two databases of responses, in which there are different values for each

sensitive variable without disturbing the underlying statistics, then a different database, which is just as usable by researchers as the first, can be released and thereby keep from compromising the confidentiality of those responses.

The example presented by Dalenius and Reiss shows a two-order data swap of one variable in a set of four binary variables. Therefore they look for a series of swaps such that all two-way marginals are left unchanged. For a set of 10 records, they swapped one variable for 4 of them. They found 5 series of such swaps involving 4 records that would preserve all 2-way marginals.

According to the Dalenius-Reiss definition of a  $k$ -order swap, all  $k$ -way marginals are preserved. No higher order marginals are guaranteed to be preserved. They present no algorithm for doing these swaps or finding which ones are available. They do, however, present theorems and statements about the probabilities of there being swaps.

### **Census Version**

The Census Bureau actually did a simulation of a variant of data swapping using 100% decennial census data from the state of New Jersey. The purpose behind the simulation was to see if this method would be acceptable as a disclosure avoidance procedure for the 1990 Census tabulations and even possibly for the release of 1990 Census microdata. (see Griffin, et al. 1989, Navarro et al., 1988, as well as Subcommittee on Disclosure Avoidance Techniques, 1994). The results were considered to be a success and essentially the same methodology was actually used for data releases from the 1990 Census.

Navarro, Flores-Baez, and Thompson (1988) describe in detail the simulation. The first part involved matching data records. The Bureau used two matching procedures. In the first procedure, two housing units would match if they matched on the following items:

1. number of persons of each race (white, black, AIAE, API, other) living in the household,
2. number of non-Hispanics living in the household,
3. number of Hispanics living in the household,
4. number of people 18 years or older living in the household,
5. number of units living at the address,
6. mobile home or trailer designation.

In the second procedure matched housing units if they had the same number of persons for every cell of the  $2 \times 2 \times 5$  three-dimensional matrix defined by

1. age (under 18 years and 18 years and over),
2. Spanish Origin and not of Spanish Origin,
3. major race (White, Black, AIAE, API, other).

and had the same number of units at the address and the same mobile home or trailer designation. The primary difference between the two procedures the cross classification of Spanish Origin by race and age. Therefore there are only 10 variables on which to match in the first procedure, whereas there are 22 variables on which to match in the second.

Using Dalenius and Reiss's terminology, the Census Bureau guaranteed that a set of specific

marginals will be the same. The pair of marginals that are guaranteed are the  $(k+1)$ -way marginals for the  $k$  variables being matched on and the location/switch variable and the  $(n-1)$ -way marginal for all variables except for the location/switch variable. But the marginals involving the location variable on which they switch and the remaining variables that are not matched on are not guaranteed to stay unchanged, even at the two-way level.

### **SAFE Version**

Appel, Kinzel, and Nölte (1993) introduce yet another method of protecting data. Their approach implicitly involves the same ideas of keeping  $k$ th-marginal totals constant and disturbing the tabulations and cross-tabulations as little as possible, but takes a more direct route towards "anonymization."

The method discovers how many times each unique combination of variable keys occurs by creating an  $n$ -dimensional contingency table. This contingency table has an implicit hierarchy of the categories for the  $n-1$  variables for columns and then uses the  $n$ th variable cross-classifying for the rows.

Thus, in other words, the number of columns equals the number of different combinations of categories for the first  $n-1$  variables and the number of rows equals the number of categories for the  $n$ th variable. Next, they select all of those combinations that appear less than 3 times in the data file. The article specifies these as being most "compromisable." Then they apply a set of five rules in a specific order to the data. Whenever any data switching takes place, it takes place within one of the columns. This means that the only tabulations that will be disturbed are those involving the cross-classified  $n$ th variable, and thus the  $(n-1)$ -way margin involving all other variables will be preserved.

The five rules are:

1. Any combination with a frequency 1 is to be counted with any combination with frequency 2.
2. If there are three combinations with frequency 1, the central combination is set equal to 3 and the other two are replaced with 0.
3. Any combination with a frequency 1 is added to the combination with the greatest frequency.
4. A frequency of 1 is subtracted from the maximum frequency and added to a combination with frequency 2.
5. A combination with frequency of 2 is split up (rarely used).

Thus, if there are any combinations of the variables that have a frequency of less than three the data will be perturbed in some fashion. The perturbations in the data will be limited to the row totals, not the column totals. Therefore to readjust the one-dimensional row statistics, one must perform some additional arbitrary adjustments on the data (referred to by the authors as *compensation*). Once these adjustments have been made so that the one-way marginals are correct and no combinations are left with frequency of less than 3, then the data are "protected" in that no single released cell will have no less than 2 observations. Also, it is only possible to try and identify those records that have been added to others by whether or not they are 0, but with a high number of possible combinations, it will be impossible to tell whether the combination had frequency 0 to start or not. This method also provides protection for those categories in which there are no keys of a certain kind, because a user

or intruder will be unable to determine whether a 0 in the tabulation is a real 0 or an adjusted 0.

The SAFE method "guarantees" the preservation of the  $(n-1)$ -way margin for the first  $n-1$  variables and the one-way margin for the  $n$ th variable as well. To do more, i.e. to preserve more margins requires some sort of iteration. Because the methods are not presented and described in the context of formal statistical models, it is unclear exactly what is optimized and whether or not the procedures can possibly converge. This relates to the notion, mentioned above for the Dalenius-Reiss method, of the "existence" of data swaps.

## Comparisons

The Census Bureau method and the SAFE method are similar in that they protect a certain set of higher-dimensional margins but are not comprehensive in the lower-dimensional margins they control. Dalenius and Reiss speak mainly of keeping fixed a complete set of  $k$ -way marginals. Thus, we can think of the Census and SAFE methods as attempting to maintain relationships at a high level for some variables, and only at a low level for others, whereas the type of data-swapping discussed originally by Dalenius and Reiss has the same "level of protection" for all variables.

Allowing for preservation of margins at different levels would seem to be important especially in the swapping of census-type data swapping, because the location variable is the one being swapped. This means that someone using the data could not be sure of the relationship between the location variable and any of the unmatched variables, even at a two-dimensional level. There is no reason why swapping needs to be restricted in this particular fashion.

In summary, the data-swapping methods described in these three different sources, when applied to purely categorical data, all have the characteristic of attempting to preserve certain pre-specified marginal totals and then moving pairs of observations from one cell to another in compensating ways. None of the papers make the link between their methods and those for the analysis of loglinear models as we do in this paper.

## ACKNOWLEDGMENTS

The preparation of this paper was supported in part under a contract with Westat and the U.S. Bureau of the Census. We are especially appreciative to David Binder and George Duncan for initial reactions to the general strategy and we thank Persi Diaconis, George Duncan, Gary Glonek, Peter Müller, Danny Pfefferman, and Steffen Lauritzen for providing references that have found their way into this paper. None of them, however, bear any responsibility for our use of the suggested materials, or for our admittedly speculative ideas on the applicability of specific statistical methods. An earlier version of material that forms the core of this paper will appear as Fienberg (1996).

## REFERENCES

Agresti, A. (1992). A survey of exact inference for contingency tables (with discussion). *Statistical Science*, **7**, 131-177.

Appel, G., Kinzel, S., and Nölte, D. (1993) SAFE-A generally usable program system for the anonymization of individual data in official statistics. *Proceedings of the International Seminar on*

*Statistical Confidentiality*, Dublin, Ireland, September 8-10, 1992, 201-228.

Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press.

Carvalho, F. de, Dellaert, N. and Osorio, M. de S. (1994). Statistical disclosure in two-dimensional tables: General tables. *Journal of the American Statistical Association*, **89**, 1547-1557.

Cox, L. (1980). Suppression methodology and statistical disclosure control. *Journal of the American Statistical Association*, **75**, 377-385.

Cox, L. (1995). Network models for complementary cell suppression. *Journal of the American Statistical Association*, **90**, 1453-1462.

Cox, L. and Sande, G. (1978). Automated statistical disclosure control. *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 177-182.

Dalenius, T. and Reiss, S.P. (1978). Data-swapping: A technique for disclosure control (extended abstract). *American Statistical Association Proceedings of the Section on Survey Research Methods*, 191-194.

Dalenius, T. and Reiss, S.P. (1982). Data-swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference*, **6**, 73-85.

Darroch, J.N., Lauritzen, S., and Speed, T.P. (1980). Markov fields and log-linear interaction models for contingency tables. *Annals of Statistics*, **8**, 522-539.

Diaconis, P. and Sturmfels, B. (1993). Algebraic algorithms for sampling from conditional distributions. Unpublished manuscript.

Duncan, G.T. and Lambert, D. (1986). Disclosure-limited data dissemination (with discussion). *Journal of American Statistical Association*, **81**, 10-28.

Duncan, G.T. and Lambert, D. (1989). The risk of disclosure for microdata. *Journal of Business and Economic Statistics*, **7**, 207-217.

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

Edwards, D. (1995). *Introduction to Graphical Modelling*. New York: Springer-Verlag.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics*, **7**, 1-26.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall.

Epstein, A.D. and Fienberg, S.E. (1992). Bayesian Estimation in multidimensional contingency tables, In *Proceedings of Indo-U. S. Workshop on Bayesian Analysis in Statistics and Econometrics*

(P.K. Goel and N.S. Iyengar, eds), Lecture Notes in Statistics Vol. 75, New York: Springer-Verlag, pp. 37-47.

Fienberg, S.E. (1975). Perspectives Canada as a social report. *Social Indicators Research*, **2**, 153-174.

Fienberg, S.E. (1980). *The Analysis of Cross-Classified Categorical Data*. (2nd ed.) Cambridge, MA: MIT Press.

Fienberg, S.E. (1994a). Conflicts between the needs for access to statistical information and demands for confidentiality. *Journal of Official Statistics*, **10**, 115-132.

Fienberg, S.E. (1994b). A radical proposal for the provision of micro-data samples and the preservation of confidentiality. Technical Report No. 611, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA.

Fienberg, S.E. (1995). Discussion of presentations on statistical disclosure methodology, In *Seminar on New Directions in Statistical Methodology, Statistical Policy Working Paper No. 23*. Federal Committee on Statistical Methodology, Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget, Washington, DC, Part 1, pp. 68-79.

Fienberg, S.E. (1996). Taking uncertainty and error in censuses and surveys seriously. *Proceedings of Statistics Canada Symposium 95: From Data to Information-Methods and Systems* (in press).

Fienberg, S.E., Makov, U.K., and Sanil, A. (1994). A Bayesian approach to data disclosure: optimal intruder behavior for continuous data. Technical Report No. 608, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA.

Fuller, W. (1993). Masking procedures for microdata disclosure. *Journal of Official Statistics*, **9**, 383-406.

Gelfand, A.E. and Mukhopadhyay, S. (1995). On Nonparametric Bayesian inference for the distribution of a random sample. *Canadian Journal of Statistics*, **23**, 411-420.

Glonek, G. (1987). *Some Aspects of Log Linear Models*. Ph.D. Thesis, School of Mathematical Sciences, Flinders University of South Australia.

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

Griffin, R., Navarro, A., and Flores-Baez, L. (1989). Disclosure avoidance for the 1990 census. *Proceedings of the Section on Survey Research*, American Statistical Association, 516-521.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. New York: Springer-Verlag.

Heer, G. R. (1993). A bootstrap procedure to preserve statistical confidentiality in contingency tables. *Proceedings of the International Seminar on Statistical Confidentiality*, Dublin, Ireland, September 8-10, 1992, 261-271.

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

Lauritzen, S. (1996). *Graphical Association Models*. Oxford University Press, New York.

Lavine, M. (1992). Some aspects of Polya tree distributions for statistical modeling. *Annals of Statistics*, **20**, 1222-1235.

Liew, C., Choi, U., and Liew C. (1985), A Data Distortion by Probability Distribution. *ACM Transactions on Database Systems*, **10**. 395-411.

Madigan, D. And York, J. (1995). Bayesian graphical models for discrete data. *International Statistical Review*, **63**, 215-232.

Meng, X-L. (1994). Multiple-imputation inferences with uncongenial sources of inputs, (with discussion). *Statistical Science*, **9**, 538-573.

Navarro, A., Flores-Baez, L., and Thompson, J. (1988). Results of Data Switching Simulation. Presented at the Spring meeting of the American Statistical Association and Population Statistics Census Advisory Committees.

Patefield, W.M. (1981). An efficient method of generating random R x C tables with given row and column totals. *Applied Statistics*, **30**, 91-95.

Rao, J.N.K. (1994). Estimation totals and distribution functions using auxiliary information at the estimation stage. *Journal of Official Statistics*, **10**, 153-165.

Robertson, D. (1993). Cell suppression at Statistics Canada. *Proceedings of the Annual Research Conference*, U. S. Bureau of the Census, U. S. Department of Commerce, Washington, DC, 107-131.

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

Rubin, D.B. (1993). Discussion, statistical disclosure limitation. *Journal of Official Statistics*, **9**, 461-468.

Scott, D.W. (1992). *Multivariate Density Estimation. Theory, Practice and Visualization*. New York: Wiley.

Subcommittee on Disclosure-Avoidance Techniques (1978). *Statistical Policy Working Paper No. 2: Report on Statistical Disclosure and Disclosure Avoidance Techniques*. Federal Committee on Statistical Methodology, Office of Federal Policy and Standards, U.S. Dept. of Commerce, Washington, DC.

Subcommittee on Disclosure-Avoidance Techniques (1994). *Statistical Policy Working*

*Paper No. 22: Report on Statistical Disclosure Limitation Methodology.* Federal Committee on Statistical Methodology, Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget, Washington, DC.

Sullivan, G. (1989). The use of added error to avoid disclosure in microdata releases. Unpublished Ph.D. dissertation, Department of Statistics, Iowa State University, Ames, Iowa.

West, M. Müller, P., and Escobar, M. (1994). Hierarchical priors and mixture models, with application in regression and density estimation. In *Aspects of Uncertainty*, (P.R. Freeman and A.F.M. Smith, eds), New York: Wiley, 363-386.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. New York: Wiley.

Zalkind, D.L. (1978). Comments on 'Data-swapping: A technique for disclosure control'. *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 195-196.

Zelterman, D., Chan, I. S.-F., and Mielke, P.W., Jr. (1995). Exact tests of significance in higher dimensional tables. *American Statistician*, **49**, 357-361.