STATISTICS IN MEDICINE
*Statist. Med.* 2001; **20**:1347–1356 (DOI: 10.1002/sim.672)

# Statistical perspectives on confidentiality and data access in public health

Stephen E. Fienberg*,†

*Department of Statistics, Carnegie Mellon University, Pittsburgh, PA, 15213-3890, U.S.A.*

SUMMARY

Confidentiality and disclosure limitation are topics that are inherently statistical but, until recently, they have received limited attention from statistical methodologists. That situation has changed considerably in the present decade. In this paper, we provide an introduction and overview of some statistical disclosure limitation issues that are of special relevance to public health studies and surveys, and the linkages to current research on bounds for multi-dimensional contingency table entries and 'simulated' categorical data. We also describe how these research methods relate to a new data access query system being developed for use by NCHS and other statistical agencies. Copyright © 2001 John Wiley & Sons, Ltd.

## 1. INTRODUCTION AND THEMES

The most important message of this paper is that confidentiality and disclosure limitation are inherently statistical issues. For far too long, they were relegated to the non-statistical part of large-scale data collection efforts and, as a consequence, the methods used to address them were *ad hoc* and conservative. Beginning with a 1977 paper by Dalenius [1] and a detailed and forward-looking 1978 report of the Subcommittee on Disclosure-Avoidance Techniques of the Federal Committee on Statistical Methodology [2], statisticians slowly began to address the issues in a systematic fashion. Twenty years later, we can look back and take stock of the growth of statistical activity and ideas in this area, (for example, by examining the recent 1994 report of the Federal Statistical Methodology Subcommittee on Disclosure-Avoidance Techniques [3]), the report of a panel of the Committee on National Statistics [4] and two special issues of the *Journal of Official Statistics*, in 1993 and 1998, as well as the proceedings of the 1998 Lisbon, Portugal Conference on Statistical Data Protection.

The traditional concerns regarding disclosure limitation and their remedies are not especially unique to public health data, whether they arise from surveys or epidemiological studies.

* Correspondence to: Stephen E. Fienberg, Department of Statistics, Carnegie Mellon University, Pittsburgh,PA, 15213-3890, U.S.A.
† E-mail: fienberg@stat.cmu.edu

Contract/grant sponsor: Westat
Contract/grant sponsor: U.S. Bureau of the Census

These methods focus on databases where the basic units of measurement and/or recording are numerical in nature, for example, the weight and height of the respondent, or codings indicating the presence or absence of specific diseases or symptoms. New issues of access and confidentiality loom on the horizon as those engaged in health research consider data elements that consist of functional magnetic resonance imaging or full body scan images, or even genetics samples (see the discussion of multiple media data in Fienberg [5]). These issues have already provoked considerable controversy in the context of data associated with stored tissue samples at CDC (for example, see Clayton *et al.* [6]). Because blood and other samples play important roles in health survey data collection, it is essential that statisticians begin thinking about how to handle their 'release', or limit their disclosure possibilities through restriction of what is released as part of a public-use data file. Such issues pose enormous methodological challenges.

In the next section, I briefly address the issue of whether we should release restricted data to achieve confidentiality objectives or whether we should simply restrict access. Both approaches have as their goal disclosure limitation. I am an advocate for unrestricted access to as much data as it is possible to release. Thus, I briefly attempt to make the case for unlimited access to restricted data so as an approach to limit disclosure risk, but not so much as to impair the vast majority of potential research uses of the data.

In Section 3, I then provide an overview of some current statistical ideas in use for data disclosure limitation and give special attention to categorical variables. One theme that comes up repeatedly in this literature is the role of bounds for multi-dimensional contingency tables. After discussing this topic in Section 4, I relate recent research ideas in disclosure limitation to a new data access query system in development for use by NCHS and other statistical agencies.

Rather than citing the extensive literature on some of these topics, I provide entry points, often via some of my recent papers on methodology for disclosure limitation. In these papers, the link between the disclosure limitation methods and more traditional statistical methodology is made explicit.

## 2. DISCLOSURE ISSUES: RESTRICTED ACCESS VERSUS RESTRICTED DATA

The probabilistic notion of disclosure, due originally to Dalenius [1], suggests that any release of actual data should produce disclosure at some level, since the released data should increase the information available about individuals in the database. This is in essence a statement about the changing conditional probability of identification of individuals as one conditions on increasing amounts of information.

There are actually two types of disclosure: exact and inferential. For exact disclosure we talk about disclosing, with probability one, the identity of an individual respondent and thus various attributes of that individual, or simply disclosure resulting from attributes possessed by a group of individuals of whom the target is one. This can happen in various ways, but more often than not we infer such identity and/or attributes, with probability less than 1. Implicit in almost all of the recent research on the topic is the role of the unidentified intruder who had data to match against the released data files (for example, see Lambert [7] and Fienberg *et al.* [8]). Thus, the intruder's goal is to effect identification and thereby create linked files.

 

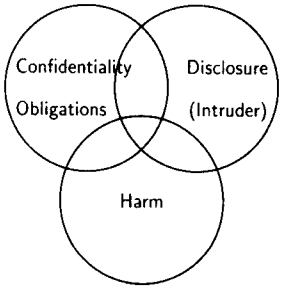


Figure 1. Relationships among confidentiality, disclosure and harm.

As Figure 1 depicts in a schematic fashion, not all data disclosures breach promises of confidentiality to respondents, since release of data for those in a sample increases the information available for those not in the sample, and an intruder can cause harm to those whose data are not released by falsely identifying someone in a database. See Fienberg [9] for further discussion along these lines.

There are two different philosophies that people adopt with regard to the preservation of confidentiality associated with individual-level data: (i) *restricted* or *limited information*, wherein the amount or format of the data released is subject to restrictions; and (ii) *restricted* or *limited access*, wherein the access to the information is itself restricted. I have argued elsewhere (for example, see Fienberg [9, 10]) that federal statistical data are a public good and that the federal statistical agencies have a *responsibility* to provide wide and unrestricted access to data that might be of value to secondary users outside the agencies themselves. Restricted access should only be justified in extreme situations where the confidentiality of data in the possession of an agency cannot be protected through some form of restriction on the information released. For a discussion of some of the benefits of restricted access, see David [11] and for a response regarding why restricted access approaches leave far too much to be desired, see Fienberg [10].

Government statistical data such as those gathered as part of censuses and major sample surveys meet two key tests that are usually applied to quantities labelled as public goods: jointness of consumption (consumption by one person does not diminish their availability to others), and benefit to the nation as a whole (statistical data are used to inform public policy and as the basis for democratic representation). The only issue, then, is whether or not there is non-exclusivity, that is, whether or not it makes sense to provide these statistical data to some citizens and not to others. If we have means for providing access to all or virtually all in society, for example, via the Internet and the World Wide Web, then the costs of providing the data to all is often less than the costs of restricting access. However there are other costs that result from expanded use to those who produce the data. For a general discussion of the costs and benefits of data sharing see Fienberg *et al.* [12] and Fienberg [13] for an updated perspective in a public health and medical context. Duncan *et al.* (reference [4], pp. 29–33) and Duncan [14] give a more focused discussion relevant to the present context.

My view is that not only is restricting access to the public a bad public policy, but that it cannot work effectively. This is primarily because the gate keepers for restricted data systems have little or no incentive to widen access or to allow research analysts the same freedom to work with a data set and share results as they are accustomed to having with unrestricted

 

Table I. Original $3 \times 2 \times 2$ table with marginals.

| | | | | | |
|---|---|---|---|---|---|
| $n_{111}$ | $n_{121}$ | $n_{1+1}$ | $n_{112}$ | $n_{122}$ | $n_{1+2}$ |
| $n_{211}$ | $n_{221}$ | $n_{2+1}$ | $n_{212}$ | $n_{222}$ | $n_{2+2}$ |
| $n_{311}$ | $n_{321}$ | $n_{3+1}$ | $n_{312}$ | $n_{322}$ | $n_{3+2}$ |
| $n_{+11}$ | $n_{+21}$ | $n_{++1}$ | $n_{+12}$ | $n_{+22}$ | $n_{++2}$ |

access. Just imagine the difficulty the researchers would have if they are accustomed to reporting residual plots and other information that allows for a partial reconstruction of the original data, at least for some variables, since restricted data centers typically do not allow users to take such information away. Thus, for me, the question is not if we should continue to supply public-use microdata, but how. For that we need tools for disclosure limitation that have as their output usable statistical data bases. That is the topic of the remainder of this paper.

## 3. PERTURBATION METHODS AND 'SIMULATED' PUBLIC-USE DATA

There is a general class of methods for disclosure limitation that were labelled *matrix masking* by Duncan and Pearson [15]. The idea is to think in terms of transforming an $n \times p$ data matrix $\mathbf{Z}$ through pre- and post-multiplication and possible addition of noise, that is

$$\mathbf{Z} \rightarrow \mathbf{AZB} + \mathbf{C} \tag{1}$$

where $\mathbf{A}$ is a matrix that operates on cases, $\mathbf{B}$ is a matrix that operates on variables, and $\mathbf{C}$ is a matrix that adds perturbations or noise. Matrix masking includes a wide variety of standard approaches to disclosure limitation:

(i) releasing a subset of observations (delete rows from $\mathbf{Z}$);
(ii) *cell suppression* for cross-classifications;
(iii) including simulated data (add rows to $\mathbf{Z}$);
(iv) releasing a subset of variables (delete columns from $\mathbf{Z}$);
(v) switching selected column values for pairs of rows. The latter approach is a technique known as *data swapping* and is described in more detail below. Data swapping was first proposed by Dalenius and Reiss [16] and now serves as the basis of the disclosure limitation methodology underpinning data releases from the U.S. decennial census.

To illustrate the method of data swapping pictorially, we consider a $3 \times 2 \times 2$ contingency table with entries $\{n_{ijk}\}$ as in Table I.

We want to track what happens when we swap the the values for a randomly selected pair of individuals, one in layer 1 and the other in layer 2. Suppose that the individual selected from layer 1 is in the $(1,2,1)$ cell and that we are swapping his/her characteristics with a randomly selected individual in the $(3,1,2)$ cell. Table II shows the result. Note that the two-dimensional total for the first two variables (adding over layers) is unchanged, as is the one-dimensional total for the third variable. This process is now repeated for pairs of randomly selected units in the two layers, thus producing a transformed table that continues to preserve the same marginal totals.

 

Table II. Altered $3 \times 2 \times 2$ table with marginals: observation from the $(1,2,1)$ cell swapped with observation from the $(3,1,2)$ cell.

| | | | | | |
|---|---|---|---|---|---|
| $n_{111}$ | $n_{121}-1$ | $n_{1+1}-1$ | $n_{112}$ | $n_{122}+1$ | $n_{1+2}+1$ |
| $n_{211}$ | $n_{221}$ | $n_{2+1}$ | $n_{212}$ | $n_{222}$ | $n_{2+2}$ |
| $n_{311}+1$ | $n_{321}$ | $n_{3+1}+1$ | $n_{312}-1$ | $n_{322}$ | $n_{3+2}-1$ |
| $n_{+11}+1$ | $n_{+21}-1$ | $n_{++1}$ | $n_{+12}-1$ | $n_{+22}+1$ | $n_{++2}$ |

Some matrix masking methods alter the data in systematic ways, for example, through aggregation or through cell suppression, and others do it through random perturbations, often subject to constraints for aggregates. Examples of perturbation methods are *controlled random rounding*, *data swapping*, and the recently proposed *post-randomization method* or PRAM of Gouweleeuw *et al.* [17] and generalized by Duncan and Fienberg [18]. One way to think about random perturbation methods is as a restricted simulation tool, and thus we can link them to other types of simulation approaches that have recently been proposed.

Fienberg *et al.* [19] pursue this simulation strategy and present a general approach to 'simulating' from a constrained version of the cumulative empirical distribution function of the data. In the case when all of the variables are categorical, the cumulative distribution function is essentially the same as the counts in the resulting cross-classification or contingency table. As a consequence, we think of this general simulation approach as equivalent to simulating from a constrained contingency table, for example, given a specific set of marginal totals and replacing the original data by a randomly generated one drawn from the 'exact' distribution of the contingency table under a log-linear model that includes 'confidentiality-preserving' margins among its minimal sufficient statistics. Actually, Fienberg *et al.* [19] propose retaining the simulated table only if it is consistent with some more complex log-linear model. This approach offers the prospect of simultaneously smoothing of the original counts *and* providing disclosure limitation protection.

How does this method relate to data swapping? The data swap transformation represented by moving from Table I to Table II actually represents one of a subclass of possible 'moves' in a Markov chain Monte Carlo algorithm proposed by Diaconis and Sturmfels [20], but such moves alone do not always suffice to generate the exact distribution. Even in the cases where they do suffice, however, one needs to run the Markov chain a very long time in order to simulate the exact distribution. Making a small proportion of swaps (as is done with the decennial census procedures) is not sufficient to provide a firm statistical foundation taking into account the added uncertainty that results from the alteration of the data.

An extremely important feature of the simulation methodology used here is that information on the variability is directly accessible to the user, since anyone can begin with the reported table and information about the margins that are held fixed, and then run the Diaconis–Sturmfels Markov chain algorithm to regenerate the full distribution of all possible tables with those margins. This then allows the user to make inference about the added variability in a formal modelling context in a form that is similar to the approach to inference in Gouweleeuw *et al.* [17]. As a consequence, simulation and perturbation methods represent a major improvement from the perspective of access to data over cell suppression and data swapping.

 

Many practical questions remain regarding the use and efficacy of such methods for generating disclosure-limited public-use samples. For example:

(i) How effective are such devices for limiting disclosure, that is, protecting against intruders?
(ii) What is information loss when we compare actual data with those released?
(iii) How can they be used when the full cross-classification of interest is very sparse, consisting largely of 0s and 1s?
(iv) How can we use models to generate the simulated data when the users have a multiplicity of models and even classes of models which they would like to apply to the released data?

Among the tools for risk assessment are various approaches for estimating the number of uniques in a population. For three quite different but none the less related approaches to this problem, see Fienberg and Makov [21], Samuels [22] and Skinner and Holmes [23]. These authors actually provide per-record assessments of risk for the categorical response case. To go with risk assessment we also need information on the trade-offs of gains that balance the risks. Few have attended to this issue. The most interesting example arises in the context of a paper by Pannekoek and de Waal [24], who suggest reporting empirical-Bayes-like mixtures of the observed data and smoothed versions of them for small area categorical data. Following their paper, Zaslavsky and Horton [25] discuss how to evaluate the trade-off between disclosure risk in their approach and the loss due to non-publication. However, much more work needs to be done on both risk assessment and its trade-offs.

## 4. BONFERRONI–FRÉCHET–HOEFFDING BOUNDS

Consider a $2 \times 2$ table of counts, $\{n_{ij}\}$, with the marginal totals $\{n_1+, n_{2+}\}$ and $\{n_{+1}, n_{+2}\}$:

| | | |
|---|---|---|
| $n_{11}$ | $n_{12}$ | $n_{1+}$ |
| $n_{21}$ | $n_{22}$ | $n_{2+}$ |
| $n_{+1}$ | $n_{+2}$ | $n$ |

The marginal constraints, that is, that the counts in any row add to the corresponding one-way total, plus the fact that the counts must be non-negative, imply bounds for the cell entries. Specifically, for the $(i,j)$ cell, we have

$$\min\{n_{i+}, n_{+j}\} \geqslant n_{ij} \geqslant \max\{n_{i+} + n_{+j} - n, 0\} \tag{2}$$

Bounds such as those in equation (2) are usually are referred to as Fréchet bounds after the French statistician Maurice Fréchet who described them in a 1940 paper (see Fréchet [26]), but they were independently described by Bonferroni and Hoeffding at about the same time. They have been repeatedly rediscovered by a myriad of others. Such bounds and their generalizations lie at the heart of a number of different approaches to disclosure limitation including cell suppression, data swapping and other random perturbation methods, and controlled rounding (for example, see the discussion by Cox [27]).

In Fienberg [28] I described these bounds and several of their multi-dimensional generalizations, especially those involving overlapping marginals. There are also fascinating statistical links here to separate literatures on the existence of maximum likelihood estimates for

 

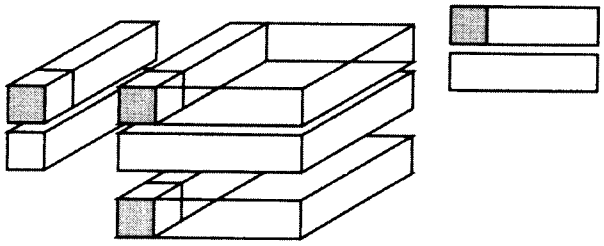

Figure 2. The marginal constraints for cells in a three-way table given two-way marginals.

log-linear models (see Bishop *et al.* [29] and Haberman [30]), Monte Carlo Markov chain techniques for generating exact condition distributions using the method of Gröbner bases (see Diaconis and Sturmfels [20] and Fienberg *et al.* [19]), and bounds for contingency tables given a set of marginals (for example, see Buzzigoli and Giusti [31], Fienberg [28] and Roehrig *et al.* [32]).

My continuing research on the topic of bounds has focused in part on the special 'graphical' sub-family of log-linear models defined in terms of simultaneous conditional independencies (for example, see Lauritzen [33]), and the subclass of 'decomposable' graphical models (called direct models by Bishop *et al.* [29]) for which simple formulae for bounds exist and are sharp. For example, consider the $k$-dimensional table with counts $\{n_{i_1 i_2 \ldots i_k}\}$, with one-way marginals given. (This corresponds to the model of complete independence among the $k$ variables.) The Fréchet bounds for the cell entries are illustrated by the following ones for the $(1,1,\ldots,1)$:

$$\min\{n_{1+\cdots+}, n_{+1+\cdots+}, \ldots, n_{+1+\cdots+},\}$$

$$\geqslant n_{11\ldots1} \geqslant \max\{n_{1+\cdots+} + n_{+1+\cdots+} + \cdots + n_{+\cdots+1} - n(k-1), 0\} \tag{3}$$

For some special cases of non-graphical models we can construct formulae for bounds, for example, for $2^k$ tables with $(k-1)$-dimensional margins fixed (we need one extra bound here to go with the natural Fréchet bounds and it comes from log-linear model theory). In Figure 2 we depict the constraining nature of such bounds for a simple three-dimensional cross-classification with fixed two-way margins.

These and other bounds play an important role in a proposed project for large scale disclosure limitation to which we now turn.

## 5. A PILOT QUERY SYSTEM FOR PUBLIC DATA ACCESS

The National Institute of Statistical Science (NISS) has recently assembled a team of statistical researchers from multiple universities who have begun to work with statisticians in U.S. statistical agencies. They are developing a Web-based query system that allows the use of disclosure limitation methods applied sequentially in response to a series of statistical queries in which the public knowledge of releases is cumulative.

The query system idea draws in part on a pilot project described in Keller-McNulty and Unger [34] and it will use as tools the various disclosure limitation methods being developed in the literature. The idea is to fully automate the methods through algorithms and

 

explore intruder behaviour (see Fienberg *et al.* [8]) and to utilize alternative approaches to risk assessment.

To get a sense of how this system *might* use the ideas on simulated databases, consider a database consisting of a $k$-dimensional contingency table, for which the queries are only allowed to come in the form of requests for marginal tables of dimension $\leqslant k-1$. What we know from statistical theory is that, as margins are released and cumulated by a user, we have increasing information about the table entries.

In response to a new query, the system now examines it in combination with all those previously released margins and decides if the bounds on the cells of the cross-classification are too tight. Then it might offer one of three responses: (i) yes – release; (ii) no – do not release; or perhaps (iii) simulate a new table, which is consistent with the previously released margins, and then release the requested margin table from it, because released margins need to be consistent and even simulated releases become highly constrained.

A sequential query system need not be restricted to categorical variables or to queries that come in the form of requests for tables. Further, my favourite simulation techniques for disclosure limitation and my work on bounds for tables will represent only one of many alternative disclosure limitation strategies that need to be explored, for example, the Argus approach developed by the statisticians at Statistics Netherlands (see Hundepool *et al.* [35, 36] and Willenborg and De Waal [37]).

NISS plans to develop a basic query system, test it with one or more public-use microdata files, test intruder behaviour in a variety of ways, and elicit agency and user reactions. While there are many theoretical and empirical issues to explore and many exciting research questions to address, making such a system function, with actual agency databases, offers the real future prospect of improved disclosure limitation *and* increased data access.

## 6. CONCLUSION

In this paper I began by explaining some of the complex relationships between promises of confidentiality to respondents in surveys or participants in studies and the nature of disclosure of information about those respondents. One cannot eliminate the risk of disclosure, simply reduce it, unless one restricts access to the data. Thus techniques for disclosure limitation are inherently statistical in nature and must be evaluated using statistical tools for assessing the risk of harm to respondents.

I presented what has become a standard framework for a large number of techniques for disclosure limitation, namely, matrix masking, and we explained how some of these techniques work. When data are categorical rather than continuous, and thus form a contingency table, most of the techniques in current use involve marginal constraints. I described a few of these including a recent proposal for simulating data for release, and linked these methods to theory associated with log-linear models. This linkage allows for stocktaking both in terms of the usefulness of the data released and the effectiveness of disclosure limitation methodologies.

The role of marginal bounds for multi-way contingency tables raises new statistical issues, and I outlined a few of these in a separate section. Then I described a project organized by the National Institute for Statistical Sciences for evaluating competing approaches using a real-time sequential query-based system. I ended by briefly outlining how such a system might work using the ideas of bounds and simulated data for contingency tables.

 