Journal of Official Statistics, Vol. 14, No. 4, 1998, pp. 385–397

Confidentiality, Uniqueness, and Disclosure Limitation for Categorical Data[1]

Stephen E. Fienberg[2] and Udi E. Makov[3]

When an agency releases data on individuals that are categorical in nature, the possible identification of those who are unique or rare in the population is a concern because identity disclosure is deemed to be a violation of promises of confidentiality. We review relationships among uniqueness in a sample, uniqueness in the population, and notions of disclosure, and then turn to methods for assessing disclosure potential as a result of sample uniqueness, especially using log-linear models.

Key words: Contingency tables; data disclosure; imputation; loglinear models; population uniqueness; sample uniqueness; exact distribution.

1. Introduction and Background

Suppose a population of individuals is cross-classified according to several categorical variables yielding a cell with an entry of ''1.'' Then we say that the individual corresponding to that ''1'' is unique in the population for these variables, or more succinctly is a *population unique*. Note that, in principle, if we use enough variables everyone in the population may be unique. Thus we presume that the data collection agency has been somewhat careful in its choice of a set of p variables to collect and the total number of cells in the resulting cross-classification, K, is sufficiently less than the population size, N, to make the problem of identifying population uniques statistically interesting.

When does the existence of a population unique lead to a data disclosure problem related to a pledge of confidentiality, e.g., not to release information collected from respondents in identifiable form? If a data release displays the information for an individual unique in the population, then an intruder will know that such an individual was included in the data base. An intruder who possesses matching data about a population unique has the potential to match his or her records against those in the data. This would lead to a formal violation of confidentiality. Further, if a subset of variables lead to uniqueness in the population then by matching records the intruder may actually learn some additional information about the unique individual beyond that already in his or her files.

[1]The work described here was supported in part by the U.S. Bureau of the Census through a contract with Westat and Carnegie Mellon University and in part by Statistics Netherlands while the first author was in residence there. We thank Chris Skinner, Leon Willenborg, and Laura Zayatz for helpful conversations while this research was in progress, and we thank a referee for several suggestions which have made their way into the final version of this article. Russell Steele assisted with critical computational support in connection with the example in Section 4. An earlier version of the article appeared as Fienberg and Makov (1996).
[2]Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213, U.S.A.
[3]Department of Statistics, Haifa University, Haifa, Israel.

Most agencies attempt to protect the confidentiality of population data by releasing (and possibly only collecting) a sample from that population. For the same p variables we have a sample cross-classification and we can identify those cells containing individuals corresponding to counts of ''1'' who are thus unique in the sample. Every person who is unique in the population is unique in the sample if selected, but being a sample unique does not necessarily mean that such an individual is also a population unique (e.g., see Bethlehem et al. 1990). Most papers in the literature attempt to work from the frequency of occurrence of sample uniques to that for population uniques, e.g., see the references in Fienberg and Makov (1996) and Chen and Keller-McNulty (1998).

Note that we may also be able to infer something about cells with population uniques from sample data and cells containing counts of ''0.'' If the intruder possesses data on a known individual whose characteristics correspond to a zero cell and if it is possible to infer that he or she is a population unique then a breach of confidentiality may take place. For instance, if this target individual is included in a future released sample, disclosure will be inevitable. Further, if we think of disclosure as a probabilistic phenomenon, it may also occur when the population contains cells with counts larger than ''1.'' For example, a count of ''2'' may allow someone with a set of almost unique characteristics to identify the only other person in the population with these characteristics. And a release of ''3'' in the absence of other knowledge, allows someone else to be linked to an intruder's data base containing these same variables, with probability of either 1/2 or 1/3 depending on whether the intruder possesses such characteristics.

In this article, we focus on using uniqueness in the sample to learn about uniqueness in the population, but same simple modifications to our approach easily allow us to explore the implications of small counts in sample cross-classifications for data disclosure. In Section 2, we outline an approach to this problem that has emerged over the past decade. In Section 3, we propose a new approach based on log-linear models for cross-classifications and imputation from the sample to the population. In Section 4, we discuss the new approach using an example based on a $3 \times 3 \times 2$ contingency table, and then, in Section 5, we explain how it can be extended to account for model uncertainty using a Bayesian model averaging approach.

2. Notation and Setup

We presume that a disclosure occurs if an intruder succeeds in linking a target individual to a microdata record and is able to verify with high probability that this link is correct and, in the case of sample data, unique. We confine our attention to the case where there is no measurement error, so that verification is straightforward. Then, once the key variable values for the target individual match those of the microdata exactly, disclosure is certain if no other individual in the population (except, perhaps, the intruder him- or herself) shares identical values for these key variables. This population uniqueness results in sample uniqueness, but the reverse is not necessarily true.

The existence of sample uniques and population uniques in actual data files naturally increases the likelihood of disclosure and hence both the agency and the intruder focus their attention on these cases, the former for disclosure limitation, the latter for achieving disclosure. An agency can take several measures to eliminate sample uniques altogether

(e.g., cell suppression or controlled rounding). These, however, reduce the applicability of the released data provided it can verify that the intruder is unlikely to establish population uniqueness (see the discussion below). Once the intruder has linked a target variable to a sample unique, he or she must assess the probability that the relevant sample unique is also a population unique before using the results of the match.

We use the following notation: N is the size of the population; M is the size of the sample held by the agency; and n is the size of the released sample. Typically, $n = M < N$, i.e., the agency releases the entire sample data, but an option which is always available is $n < M < N$, i.e., the agency releases only part of the collected data. Let K be the maximum variety of individuals as defined by the key variables. Since we assume that the variables are categorical, K is equal to the number of cells in the corresponding multiway contingency table, with cell probabilities, π_i, $i = l, 2, \ldots, K$. Finally, we let F_i and f_i, $i = 1, \ldots, K$, denote the counts in the cells of a multiway table summarizing the entire population and the sample, respectively.

The evaluation of $P(F_i|f_i)$ is of cardinal importance for both the agency and the intruder since

$$\sum_{i=1}^{K} P(F_i = 1|f_i = 1) \tag{1}$$

is a crucial measure of the vulnerability of the released data. In the spirit of the previous section, we can extend Equation 1 to refer to a broader definition of uniqueness such as

$$\sum_{i} [P(F_i \in [1, 2]|f_i = 1) + P(F_i = 2|f_i = 2)] \tag{2}$$

Most prior attempts to model $P(F_i|f_i)$ or to estimate the number of population uniques are based on the assumption that the probabilities associated with cell frequencies in the population are a realization from a superpopulation, e.g.,

$$\begin{aligned} F_i|\pi_i &\sim \text{Poisson}(N\pi_i) \\ f_i|\pi_i &\sim \text{Poisson}(n\pi_i) \\ n\pi_i &\sim \text{Gamma}(\alpha, \beta) \end{aligned} \tag{3}$$

As Bethlehem et al. (1990), Skinner et al.(1994), and others show, this structure results in

$$P(\text{population unique}) = (1 + N\beta)^{-(1+\alpha)} \tag{4}$$

The underlying assumption here is that the $\{\pi_i\}$ are the realization of a single density, i.e., the $\{\pi_i\}$ are assumed to be exchangeable and thus they do not reflect any underlying structure that is at the root of the model that generated the data. This approach is often referred to as one based on the *frequency of frequencies* (Good, 1953; Bishop, Fienberg, and Holland 1975, Chapter 9). Indeed, estimating the extent of population uniques in real data commonly resulted in unsatisfactory results, e.g., see Bethlehem, Keller, and Pannekoek (1990). This led Skinner et al. (1994, p. 48) to conclude ''it seems therefore that the Poisson-gamma model itself must be questioned.''

Skinner and Holmes (1998) also argue against the use of (3), because it is constant across records which are sample uniques. Instead they suggest modeling π_i as a log-linear

model, having a regression structure which reflects the particular characteristics of the ith cell:

$$\log(\pi_i) = g(\mathbf{u}_i) + \varepsilon \tag{5}$$

where $\varepsilon \sim N(0, \sigma^2)$ and the $\mathbf{u}_i$'s are loglinear model parameters similar to those in the model we work with below. The $\mathbf{u}_i$'s and σ are unknown parameters.

Samuels (1998) proposes an innovative alternative approach based on statistical estimation in the context of species sampling and population genetics, and he reanalyzed a data set from Chen and Keller-McNulty (1998).

3. Proposed Approach

Suppose that a released sample of n observations is drawn from a population of size N with cell probabilities $\{\pi_i^{(N)}\}$. Suppose further that the $\{\pi_i^{(N)}\}$ follow a log-linear model including various terms such as main effects interactions (see Bishop et al. 1975 or Fienberg 1980) of the form

$$\log(\pi_i^{(N)}) = g_N(\mathbf{u}_i) \tag{6}$$

Given a sample of size n, we propose to select a log-linear model, fit it to the observed counts $\{f_i\}$, and thus produce estimated cell probabilities $\{\hat{\pi}_i^{(n)}\}$. These may differ from $\{\pi_i^{(N)}\}$ because the fitted model differs from (6) or because the estimates of the parameters in (6) differ from their population values.

We proceed *as if* we have reasonably precise estimates of the $\{\pi_i^{(N)}\}$ based on the sample data, i.e., $\{\hat{\pi}_i^{(n)}\}$ differs from $\{\pi_i^{(N)}\}$ only as a consequence of the sampling error associated with the estimates of the margins corresponding to the highest order terms in the model of Expression 6. The estimated probabilities of the cell counts are no longer exchangeable.

Given the formal statistical model-based approach which we have pursued to this point, one might try to develop analytical formulae for the estimation of $P(F_i = 1 | f_i = 1)$. Because many of the loglinear models that might result from the estimation process will not have closed-form representations in terms of the marginal minimal sufficient statistics (see, e.g., Bishop et al. 1975), at best we would have to do so using some form of iteration or analytical approximation (cf. Skinner and Holmes 1998). Instead we have opted for a simulation-like approach.

To estimate $P(F_i = 1 | f_i = 1)$, we propose the following. Use the records on the n individuals in the sample $x_1, \ldots, x_n$ to generate from $\{\hat{\pi}_i^{(n)}\}$ $(N - n) \times H$ records, resulting in H populations of size N, where in each $(N - n)$ ''new'' records are obtained by some form of imputation (e.g., see Little and Rubin l987) or multiple imputation from a posterior distribution (e.g., see Rubin l987). Thus we have H sets of records of the form:

$$\begin{array}{llll} x_1, \ldots, x_n, & x_{n+1,1}, \ldots, x_{N,1} \\ \quad \ldots & \\ x_1, \ldots, x_n, & x_{n+1,H}, \ldots, x_{N,H} \end{array}$$

There are many ways to generate the H replicates or ''populations.'' One possibility is to use the estimated probabilities, $\{\hat{\pi}_i^{(n)}\}$, to draw a series of H independent multinomial samples of size $(N - n)$, proceeding as if the $\{\hat{\pi}_i^{(n)}\}$ are the correct cell probabilities. If we

formally include model uncertainty in a Bayesian approach, the replicates become a version of multiple imputation. In another approach, demonstrated in the next section, we can generate the imputed records from the exact distribution associated with the released contingency table given a set of fixed margins that are the minimal sufficient statistics of a log linear model.

Let $\bar{F}_i(j) = \bar{F}_i(x_1, \ldots, x_{N,j})$ be the count in cell i, based on the jth imputed population. Similarly, let $\bar{f}_i = \bar{f}_i(x_1, \ldots, x_n)$ be the count of the ith cell given the released data. Clearly $\bar{f}_i \neq 1 \Rightarrow \bar{F}_i(j) \neq 1$. What is of interest, however, is whether

$$(\bar{F}_i(j) = 1) \cap (\bar{f}_i = 1) \neq 0 \tag{7}$$

We can estimate P(population unique $-$ sample unique) by

$$\sum_{i=1}^{K} \hat{P}(F_i = 1 | f_i = 1) = \sum_{i=1}^{K} \sum_{j=1}^{H} \frac{\delta_{(\bar{F}_i(j)=1)\cap(\bar{f}_i=1)\neq 0}}{H} \tag{8}$$

where $\delta_a = 1$ if condition a is met, and $\delta_a = 0$, otherwise.

The intruder can evaluate Equation (8) and, if this estimated probability proves to be very small, should conclude that disclosure is unlikely. Actually, a disclosure takes place if the following occurs: (A) the records of the target individual are released by the agency; (B) the record of the target individual constitutes a sample unique; and (C) this sample unique is also a population unique. The probability of disclosure is equal to $P(A)P(B|A)P(C|A, B)$ and this needs to be evaluated (e.g., see Skinner et al. 1994). Clearly, an agency has to control the size of this product by releasing a small sample or by taking other measures to reduce the likelihood of identifying population uniques and, ultimately, by perturbing the released records.

Since Equation 8 is likely to decrease as $(N - n)$ increases, the agency is motivated to reduce n such that Equation 8 indicates that disclosure is infeasible. i.e., when considering the release of a sample of size n, the agency can impute the population and evaluate Equation (8) as a means of assessing the vulnerability of the released data. To aid in this assessment process we might want an estimate of the standard error for the estimated probabilities associated with Equation 8. This estimate would need to come from the replicates of the exact distribution used in the simulation process, but we do not attempt to estimate it here. What seems clear, however, is that this standard error should be smaller than those that would come from other, less-constrained simulation approaches. Finally, the agency also must weigh the tradeoffs between increased confidentiality protection and the increased uncertainty in the released data.

Actually, if we return to Equation 8 and remove the summation over the cells as indexed by the subscript i, we obtain a cell-specific measure of disclose risk. This is, in fact, the approach we adopt in the following example.

4. Example

In this section we apply an approach of imputing the ''missing records'' by generating the exact distribution associated with a contingency table given a set of fixed margins that are the minimal sufficient statistics of a loglinear model. The methods are based on an approach in Diaconis and Sturmfels (1998) and implemented and described in further

Table 1. Three-way cross-classification of gender, race, and income for a selected U.S. census tract. (Source: 1990 Census Public Use Microdata Files)

	Gender = Male Income level			
Race	≤ $10,000	> $10,000 and ≤ $25,000	> $25,000	Total
White	96	72	161	329
Black	10	7	6	23
Chinese	1[a]	1[a]	2[b]	4
Total	107	80	169	356
	Gender = Female Income level			
Race	≤ $10,000	> $10,000 and ≤ $25,000	> $25,000	Total
White	186	127	51	364
Black	11	7	3	21
Chinese	0	1[a]	0	1
Total	197	135	54	386

Note: [a]entries correspond to sample uniques; [b]entry corresponds to a ''near'' unique.

detail by Fienberg et al. (1997). Appendix 1 provides further simulation details. The data, reproduced here in Table l, form a $3 \times 3 \times 2$ table giving counts for Race by Income Group by Gender.

We ran a simulation based on draws from the exact distribution of the data in Table 1 under the no 2nd-order interaction model. No simpler loglinear model appears appropriate for this example, thus simplifying the complex problem of model selection and its implications. Our goal was to estimate the probability of a unique cell in the population given a unique cell in the sample. Using the released table as a base set of cell counts, we generated 500 tables before picking a table and adding its cell counts to the original table. The resulting table could be regarded as a population table if our sample was a 50% sample of the population. (We walked through 500 tables in order to reduce potential correlation.) In a similar way we then repeatedly selected tables at 500 table intervals during the random walk to create similar population tables for sample sizes 20%, 10%, and 5%. We obtained 100,000 of these tables for each of the four sampling fractions.

Table 2. Number of uniques in the population (100,000)

cell	count	sample fraction			
		50%	20%	10%	5%
(3,1,1)	1	31,098	929	3	0
	2	68,902	8,330	56	0
(3,2,1)	1	0	0	0	0
	2	58,234	0	0	0
(3,3,1)	1	0	0	0	0
	2	0	0	0	0
(3,2,2)	1	41,766	3,108	43	0
	2	58,234	17,184	493	0

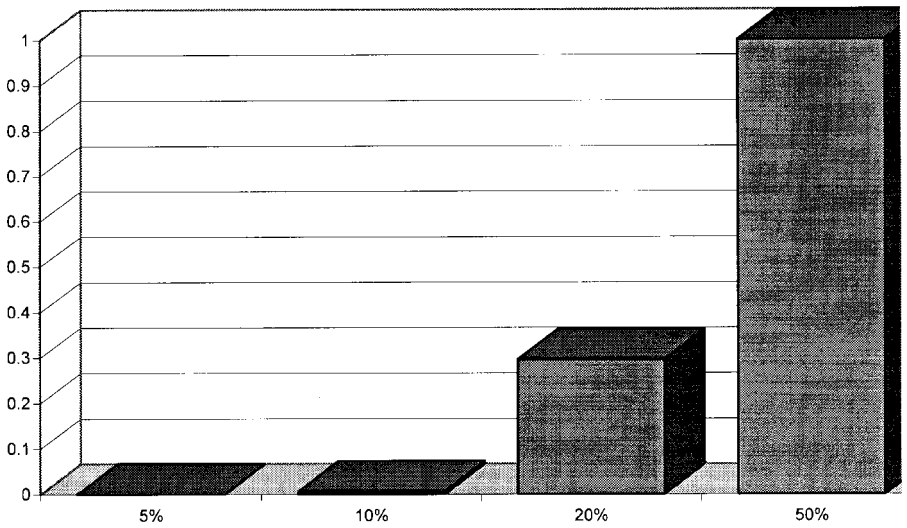

Fig. 1. $\hat{P}(C|A,B)$ *for different sample fractions*

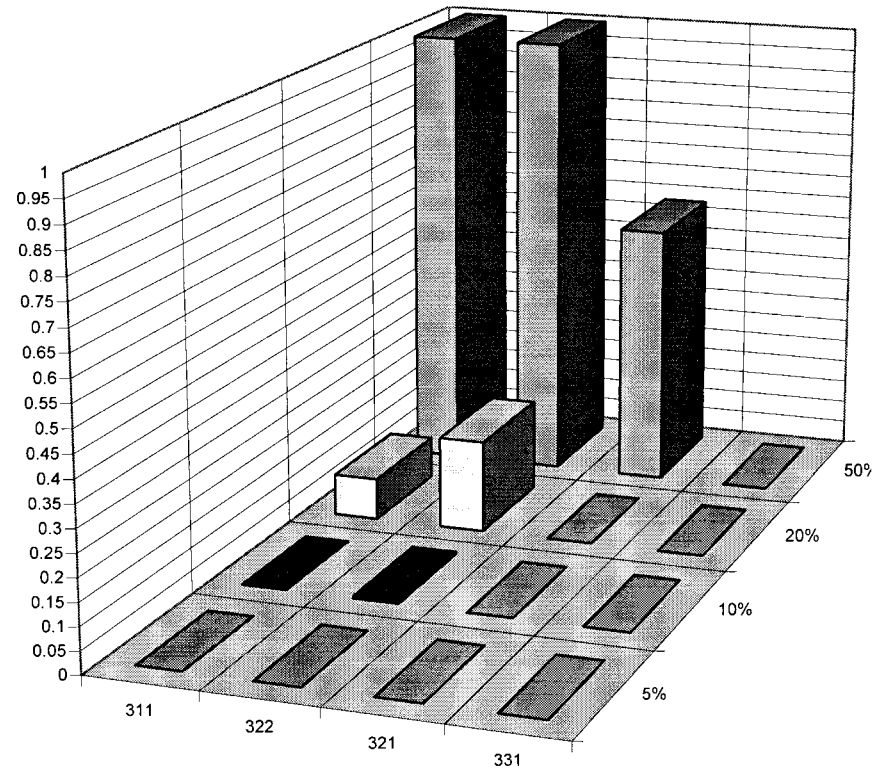

Fig. 2. $\hat{P}(C|A,B)$ *for different sample fractions, by cell*

Table 3. $\hat{P}(C|A,B)$ for zero cells in the released sample

cell	sample fraction			
	50%	20%	10%	5%
(3,1,2)	0.311	0.0682	0.0398	0.036
(3,3,2)	0.107	0.369	0.578	0.55

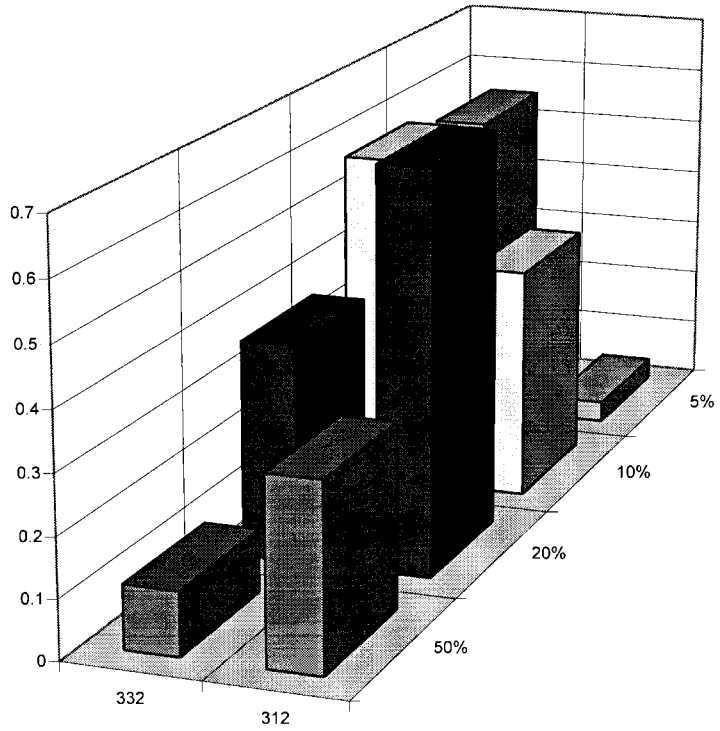

Fig. 3. Estimated probabilities of population uniques for zero cells

5. Discussion and Extensions

The method proposed in this article is linked to the fitting and estimation of a loglinear model to counts in a multi-way contingency table and it does not directly account for possible model mis-specification or added uncertainty associated with model choice or selection. The Bayesian framework allows a natural way to do this via an approach called model averaging and we outline here elements of the relevant methodology. We follow the general strategies described in Clyde (1998).

Suppose we need to consider Q possible hierarchical loglinear models, which we denote by $M_1, M_2, \ldots, M_Q$. Then the posterior distribution of disclosure (in the sense of Equation 1) is given by

$$\sum_{j=1}^{Q} \sum_{i=1}^{K} P(F_i = 1 | f_i = 1, D) P(M_j | D) \tag{9}$$

which is an average of the posterior probability of population uniques under each of the models, weighted by the respective posterior model probabilities. Here D is the available data and $P(M_j|D)$ is the posterior probability of the jth model, given by

$$P(M_j|D) = \frac{P(D|M_j)P(M_j)}{\sum_{j=1}^{Q} P(D|M_j)P(M_j)} \tag{10}$$

where

$$P(D|M_j) = \int P(D|\theta_j, M_j)P(\theta_j|M_j)d\theta_j \tag{11}$$

is the marginal likelihood of model M_j, θ_j is a vector of parameters characterizing the jth model, $P(\theta_j|M_j)$ is the prior distribution of θ_j and $P(M_j)$ is the prior probability of model j. This formulation requires the specification of θ_j which is missing from this article as we adopted an alternative approach based on the exact distribution. For Bayesian inference in multidimensional contingency tables, where θ is incorporated into the analysis, see Epstein and Fienberg (1991).

A major difficulty in implementing Bayesian model averaging arises when the number of potential models Q becomes too excessive. Clyde (1998) outlines two possible approaches: a deterministic search suggested by Madigan and Raftery (1994) and a non-deterministic search based on Markov chain Monte Carlo (MCMC) methods. In particular, a reversible jump MCMC can be adopted in which sampling is carried out from the posterior distributions of both parameters $\{\theta_j\}$ and models $\{M_j\}$. See Clyde (1998) for further information and relevant references.

In our example, fortunately, we did not have a major set of issues to explore regarding modelling error. There are only eight possible unsaturated hierarchical models for a three-way table, and in our example none of the other seven models seemed at all appropriate. We expect that had we implemented the program just outlined, we would have ended up with most of the posterior probability on the no 2^{nd}-order interaction model and the saturated model.

We plan to implement the approach outlined here for k-way tables with $k \geq 4$ in the near future and apply the methodology to an actual contingency table for which there is interest in disclosure limitation.

Appendix 1: Simulation Methodology Details

The basic algorithm used for generating possible populations from a given sample is as follows:

Read table, moves, list of interesting cells, population size indices
into respective structures
for **i** := 1 to (**number of iterations** × 500 × **max(1/X)** − 1)
total table := original table
for **j** := 1 to **max(1/X)** − 1
ctr := 0
while **ctr** < 500
r1 := *randomly generated number from* 1 *to number of moves*

```
            temporarily make move r1 to find table probability
            r2 := randomly generated number from Unif[0,1]
            if table probability > r2 and move creates no negative cells
                then make the move permanently
               else do not make the move permanently
            ctr := ctr + 1
          end while
          total table := total table + current table
          if j is in population size index list
              print interesting cells from the current table
      next j
  next i
```

This algorithm is a little more complex than the one described in the appendix to Fienberg, Meyer, Steele, and Makov (1997). Using the same logic as they use and following Diaconis and Sturmfels (1998), for each table we want to generate, we need to generate 500 so that they will be relatively uncorrelated. For each set of populations that we want to generate (a set includes one population apiece for the 5%, 10%, 20%, and 50% samples) we need to generate $1/X - 1$ tables where X equals the sample proportion we are working with. But, if we reuse the population we generate for a 50% sample when making the 20% sample, we can reduce the number of tables we need to generate for a single iteration to $500 \times$ the maximum $(1/X)$ over all X that we want to generate. In order to save space, the algorithm also permits users to select cells to display at all levels of sampling. Note that the main part of the algorithm that handles the theory mentioned in Diaconis and Sturmfels is exactly the same as the main part of the first algorithm.

Appendix 2: Detailed Simulation Results

Table 4. Estimating population distribution for cells with small sample entries, for different sampling fractions. (The first entry is the population cell count and the second is the frequency obtained in 100,000 imputations)

Cell (3,1,1)

50%	1	2								
	31,098	68,902								
20%	1	2	3	4	5					
	929	8,330	27,656	40,569	22,516					
10%	1	2	3	4	5	6	7	8	9	10
	3	56	485	2,572	8,334	18,330	26,973	25,675	14,091	3,481
5%	5	6	7	8	9	10	11	12	13	14
	16	60	327	1,023	2,827	6,035	11,195	16,274	19,548	18,374
	15	16	17	18	19	20				
	1	13,288	7,351	2,895	701	85				

Cell (3,2,1)

50%	2	3								
	58,234	41,766								
20%	5	6	7	8	9					
	11,351	32,727	35,630	17,184	3,108					
10%	10	11	12	13	14	15	16	17	18	19
	792	4,844	13,972	23,622	25,891	18,560	8,979	2,804	493	43
5%	20	21	22	23	24	25	26	27	28	29
	4	60	291	1,164	3,432	7,386	12,569	16,651	18,289	16,273
	30	31	32	33	34	35	36			
	11,606	7,095	3,470	1,241	364	93	12			

Cell (3,3,1)

50%	3	4								
	10,668	89,332								
20%	6	7	8	9	10					
	17	439	5,604	30,791	63,149					
10%	13	14	15	16	17	18	19	20		
	1	6	100	1,005	5,349	18,828	38,933	35,778		
5%	31	32	33	34	35	36	37	38	39	40
	3	45	219	956	3,493	9,496	19,441	28,586	26,380	11,381

Cell (3,1,2)

50%	0	1								
	68,902	31,098								
20%	0	1	2	3	4					
	22,516	40,569	27,656	8,330	929					
10%	0	1	2	3	4	5	6	7	8	9
	3,481	14,091	25,675	26,973	18,330	8,334	2,572	485	56	3
5%	0	1	2	3	4	5	6	7	8	9
	85	701	2,895	7,351	13,288	18,374	19,548	16,274	11,195	6,035
	10	11	12	13	14	15				
	2,827	1,023	327	60	16	1				

Cell (3,2,2)

50%	1	2								
	41,766	58,234								
20%	1	2	3	4	5					
	3,108	17,184	35,630	32,727	11,351					
10%	1	2	3	4	5	6	7	8	9	10
	43	493	2,804	8,979	18,560	25,891	23,622	13,972	4,844	792
5%	4	5	6	7	8	9	10	11	12	13
	12	93	364	1,241	3,470	7,095	11,606	16,273	18,289	16,651
	14	15	16	17	18	19	20			
	12,569	7,386	3,432	1,164	291	60	4			

Cell (3,2,3)

50%	0	1								
	89,332	10,668								
20%	0	1	2	3	4					
	63,149	30,791	5,604	439	17					
10%	0	1	2	3	4	5	6	7		
	35,778	38,933	18,828	5,349	1,005	100	6	1		
5%	0	1	2	3	4	5	6	7	8	9
	11,381	26,380	28,586	19,441	9,496	3,493	956	219	45	3

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.

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

Chen, G. and Keller-McNulty, S. (1998). Estimation of Identification Disclosure Risk in Microdata. Journal of Official Statistics, 14, 79–95.

Clyde, M.A. (1998). Bayesian Model Averaging and Model Search Strategies. In J. Bernardo et al. (Eds) Bayesian Statistics 6, Oxford University Press, Oxford, U.K.

Diaconis, P. and Sturmfels, B. (1998). Algebraic Algorithms for Sampling from Conditional Distributions. Annals of Statistics, 26, 363–397.

Epstein, A. and Fienberg, S.E. (1991). Using Gibbs Sampling for Bayesian Inference in Multidimensional Contingency Tables. Computing Science and Statistics, Interface 1991 (E.M. Keramidis (ed.)). Interface Foundation of America, 215–223.

Fienberg, S.E. (1980). The Analysis of Cross-classified Categorical Data. MIT Press, Cambridge, MA.

Fienberg, S.E. and Makov, U.E. (1996). Confidentiality, Uniqueness, and Disclosure Avoidance for Categorical Data. Proceedings of the 3rd International Seminar on Confidentiality. Bled, Slovenia, 165–174.

Fienberg, S.E., Meyer, M.M., Steele, R.J., and Makov, U.E. (1997). Notes on Generating the Exact Distribution for a Contingency Table Given Its Marginal Totals. Unpublished manuscript.

Good, I.J. (1953). On the Population Frequencies of Species and the Estimation of Population Parameters. Biometrika, 40, 237–264.

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

Madigan, D.M. and Raftery, A.E. (1994). Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window. Journal of the American Statistical Association, 89, 1535–1546.

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

Samuels, S.M. (1998) . A Bayesian, Species-sampling-inspired Approach to the Uniques Problem in Microdata Disclosure Risk Assessment. Journal of Official Statistics, 14, 373–383.

Skinner, C.J. and Holmes, D.J. (1998). Estimating the Re-identification Risk Per Record in Microdata. Journal of Official Statistics, 14, 361–372.

Skinner, C.J., Marsh, C., Openshaw, S., and Wymer, C. (1994). Disclosure Control for Census Microdata. Journal of Official Statistics, 10 , 31–51.

Received September 1997
Revised September 1998