Journal of Official Statistics, Vol. 14, No. 1, 1998, pp. 79–95

Estimation of Identification Disclosure Risk in Microdata

Guang Chen[1] *and Sallie Keller-McNulty*[2]

The necessary condition for the occurrence of an identification disclosure is that a target entity, i.e., population unit, can be uniquely identified by some set of characteristics in the population. Therefore, the percentage of the records in a released data set which can be uniquely identified in the population is an important measure of identification disclosure risk in microdata. This research deals with the development of a technique to estimate the number of unique entities in a population based on sample information. A few estimation methods have been developed for this problem, but none of them work well for small sampling fractions. To improve on the existing methods, a model for population cell frequency distributions is developed. A sample cell frequency distribution is derived assuming binomial sampling from each cell. The estimation method of model parameters based on sample cell frequencies is given. This estimation technique is tested intensively on real and simulated data sets. The results show a remarkable improvement over the existing methods, especially for sampling fractions less than 0.1.

Key words: Data security; disclosure risk; population uniqueness.

1. Introduction

Data are frequently collected and released by private businesses and government agencies for the purpose of statistical analysis. These data are typically collected with the assurance that data, on the micro level, will be kept confidential. The concern for maintaining data confidentiality coupled with the lack of methods to assess disclosure risk severely limits the ability of data collectors to provide data for legitimate research uses.

A considerable amount of research has been done in the data confidentiality area. The research addresses the ways in which the data are released to users, the methods some users might use to attack the released data, the techniques which can be applied to protect data confidentiality, and appropriate measures for disclosure risk (Adam and Wortmann 1989; Blien, Wirth, and Müller 1992; Denning and Denning 1979; Denning, Denning, and Schwartz 1979; Duncan and Lambert 1986 and 1989; Duncan and Pearson 1991; Fuller 1993; Greenberg and Zayatz 1992; Lambert 1993; Keller and Bethlehem 1992; Keller-McNulty and Unger 1993; Paass 1988; Skinner 1992; Skinner and Holmes 1992; Skinner, Marsh, Openshaw, and Wymer 1994; Spruill 1983). A common conclusion in this research is that the risk of disclosure cannot be eliminated completely. It can only

[1] Pacific Northwest Laboratory, P.O. Box 999, MS K-12, Richland, WA 99352, U.S.A.
[2] Department of Statistics, Kansas State University, Manhattan, KS 66506, U.S.A.

Acknowledgments: This research was partially supported by U.S. Bureau of the Census JSA 91-26. The authors are grateful to the referees and the Associate Editor for their helpful comments.

be maintained at a certain acceptable level. Therefore, it is very important to find methods to assess the disclosure risk for released data.

Different types of disclosure have been discussed in the literature (Duncan and Lambert 1989; Skinner 1992; Lambert 1993). The focus of this article is on *identification disclosure*. Identification disclosure occurs when a one-to-one relationship can be established between a record in the released data and a specific entity, (e.g., population unit). Implicit in this definition is the fact that the entity must be represented in the released data.

A common characteristic evident in the literature is that for given pre-knowledge, identification disclosure can occur only if a data record has a set of unique characteristic variable values, or *key variable* values, in the released data and also the corresponding entity has unique key variable values in the population. The entities with unique key variable values in a population are called *population uniques*.

The number or percentage of population uniques in the released data plays an important role in the measurement of identification disclosure risk (Bethlehem, Keller, and Pannekoek 1990). However, this measurement is not easy to obtain. In practice, the collected data are frequently a random sample from some corresponding population. Some records in the collected data may have unique key variable values. These records are called *sample uniques*. A sample unique may or may not be a population unique. Without population information, it is impossible to be certain which sample uniques are truly population uniques. Therefore, the number of the population uniques that may be in the collected data needs to be estimated based on the sample information.

This research deals with the development of a new technique to estimate the number of unique entities in a population based on information contained in a sample. A few estimation techniques have been developed in the past. These will be briefly discussed in Section 3. It is the case that none of these methods work well for small sampling fractions. This may be due to flaws in their basic model assumptions. The new method is developed in Section 4. It shows remarkable improvement over existing methods, especially for small sampling fractions. Section 5 provides discussions on variance estimation for this problem. The next section gives a description of five examples that will be used throughout this article to demonstrate the performance of the estimators. These represent a small sample of many real and simulated populations that have been studied in connection with this research.

2. Description of Example Data Sets

Three real population data sets and two simulated data sets will be used for demonstration purposes throughout this article. The five data sets are briefly described in this section.

Examples 1 and 2 are real population data sets. They represent a complete census from a single geographic region taken during the 1980 decennial census (Zayatz 1991a, 1991b). Both data sets contain the same 87,959 household records. Example 1 has five key variables: number of children, disability, employment status, marital status, and veteran status. The total number of non-empty cells defined by cross-classifying these five variables is 1,024. Among them, 222 cells have size one, which is the number of population uniques in this data set. Example 2 has seven key variables: disability, marital status, rent or mortgage payment, race, social security, tenure, and veteran status.

Cross-classifying the seven key variables generated 6,573 cells with a positive count. There are 3,105 population uniques in Example 2.

Example 3 is also a real population data set. It is a complete university administrative student database containing 6,270 undergraduate student records. The five variables: sex, marital status, ethnicity, citizenship, and curriculum were used as key variables in this data set. Cross-classifying these variables resulted in a total of 558 cells with a positive count and 278 cells of size one. This data set is used as an example of populations with small size.

The data for Example 4 were simulated to illustrate a population of a larger size. The data set contains 500,000 records and is defined by six key variables. It is not clear what a ''real'' population data set should look like. Many relationships among the distributions of the key variables could exist. To put some reasonable structure on the data, a six-dimensional multivariate normal distribution was used with the covariance structure given in Kendall (1980). The covariance structure was for 15 discrete scales of human response. The covariance structure among the first six variables was used in the simulation. It covers a range of correlation values from 0.0 to 0.8. Once the six-dimensional multivariate normal data were generated, they were discretized by dividing each marginal range into several equally spaced intervals, one for each category. Cross-classifying the six variables generated 15,047 cells with a positive count and 4,479 population uniques.

The data for Example 5 were simulated to illustrate an extreme population distribution. All five variables used to generate this data set are mutually independent with uniform marginal distributions. Therefore, the joint distribution is also a uniform distribution, which is rarely the case in reality. This data set was only used to explore the behavior of the new estimation method developed in this article. One feature that distinguishes this data set from others is that its population cell size frequency distribution has a very short tail. There are 100,000 records and 65,640 non-empty cells in the data set. The number of population uniques is 40,217.

3. Existing Methods

The structure of the population unique estimation problem can be described as follows. Suppose a population of size N is partitioned into K cells by a set of key variables, where K denotes the number of cells with a positive count. A simple random sample of size n is taken from the population. This random sample is partitioned into $k \leq K$ cells. Typically, not all population cells will be observed. The basic problem is to determine the number of cells of size one in the sample which also have size one in the population. This would represent the number of unique population entities in the data to be released.

A cell with a single entity in a sample need not correspond to a cell with a single entity in the population. Usually it is not possible to determine which cells among sample uniques are truly uniques in the population. Therefore, the number of population uniques in the sample must be estimated by using the sample information.

Several procedures have been developed in the past for the estimation of population uniqueness. Among them two procedures are widely discussed in the literature: an equivalence class procedure (Zayatz 1991a, 1991b; Greenberg and Zayatz 1992) and a Poisson-Gamma model (Bethlehem, Keller, and Pannekoek 1990; Keller and Bethlehem 1992; Skinner et al. 1994). These procedures will be discussed briefly below.

3.1. Equivalence class procedure

The equivalence class (EQC) procedure (Zayatz 1991a, 1991b; Greenberg and Zayatz 1992) is based on Bayes' rule. An equivalence class is a non-empty cell. The size of an equivalence class is the cell size. According to the Bayes' rule, the conditional probability that an observed equivalence class of size one in the sample came from a population equivalence class of size one, that is, the probability that an observed sample unique is also a population unique can be written as

$$P(1_p|1_s) = \frac{p_1 P(1_s|1_p)}{\sum_{allj} p_j P(1_s|j)}$$

where p_j is the proportion of equivalence classes of size j in the population, and $P(1_s|j)$ follows a hypergeometric distribution for all j's. The total number of population uniques, U_p, can be estimated based on the estimate of this probability. The population proportions p_j's are estimated by the sample proportions c_j/k, for all j's; where c_j is the number of cells of size j in the sample.

This procedure seems to work reasonably well for large sampling fractions, i.e., $f \geq 0.1$. However, for small sampling fractions this procedure can grossly overestimate the number of population uniques. Table 1 summarizes the estimated values for U_p based on random samples selected from the population data sets of Examples 1, 2, and 3. For each sampling fraction, 1,000 random samples were selected. Table 1 gives the average (avg) and standard deviation (sd) among the 1,000 estimates of U_p.

The equivalence class procedure clearly behaves differently for different sampling fractions. This is possibly due to the estimators used for the p_j's. With a small sampling fraction, the sample proportion structure of the equivalence classes may not correspond to the population proportion structure, thus causing the overestimation of U_p as indicated in Table 1. Note, however, the equivalence class procedure is a consistent estimation procedure in the sense that the estimate equals U_p when the entire population is sampled.

3.2. Poisson-Gamma model

Bethlehem, Keller, and Pannekoek (1990), Keller and Bethlehem (1992), and Skinner et al. (1994) proposed a Poisson-Gamma (P-G) model for the estimation of population uniqueness. This model assumes that the cell size structure in the population is a realization

Table 1. Averages and standard deviations of $\hat{U}_p$ by EQC method based on 1,000 samples

	Example 1		Example 2		Example 3	
	$U_p = 222$		$U_p = 3,105$		$U_p = 278$	
f	avg	sd	avg	sd	avg	sd
0.01	1,816	325.66	13,207	1,214.87	1,591	469.53
0.05	530	70.57	5,643	329.82	368	79.58
0.1	349	42.74	4,028	191.32	274	45.48
0.5	232	19.88	3,098	72.63	270	20.88
1.0	222	0.0	3,105	0.0	278	0.0

from a superpopulation distribution. They assume that cell sizes, denoted by Y_i, for $i = 1, 2, \ldots, K$, in the population are independent Poisson random variables with $E(Y_i) = N\pi_i$ and that the π_i's are independent and identically distributed Gamma random variables with parameters α and β. The marginal distribution of Y_i is then a negative binomial distribution with parameters α and $1 + N\beta$. Under this model, the expected number of population uniques is

$$U_p = KP(Y_1 = 1) = KN\alpha\beta(1 + N\beta)^{-(1+\alpha)}$$

The parameters, α and β, are estimated by method of moments by assuming that the sample cell frequency structure is a representative realization of the population cell frequency structure. The value of K is assumed known for this procedure.

The estimates for U_p in Table 2 were obtained by applying this model to 1,000 random samples for each sampling fraction from the data sets of Examples 1, 2, and 3. The parameters α and β were estimated using a direct application of method of moments on a negative binomial distribution with parameters α and $1 + N\beta$. This results in $\hat{\beta} = (s^2/\bar{c} - 1)/N$ and $\hat{\alpha} = \bar{c}/(N\hat{\beta})$, where $\bar{c}$ and s^2 are the mean and standard deviation of the sample cell frequencies, respectively. By further imposing the condition $K\alpha\beta = 1$, the estimator of U_p becomes $\hat{U}_p = N(1 + N\hat{\beta})^{-(1+\hat{\alpha})}$. The results show overestimation for small sampling fractions and underestimation as the sampling fraction increases. Severe underestimation was observed with a sampling fraction of 1, i.e., the entire population. This indicates that the estimation procedure does not consistently reproduce U_p when the entire population is sampled. Alternative parameterizations were also suggested. However, the empirical results on the example data showed that those alternatives would produce even more severe underestimation. Greenberg and Zayatz (1992) applied this model to the prediction of the number of population uniques in nine different U.S. Census Bureau data sets and found the estimator performed similarly as shown in Table 2.

An alternative estimator for the P-G model was applied to data from the Italian Census Bureau by Skinner et al. (1994). In this modified version, the cell size distribution in a sample is adjusted by the sample size, n. The cell sizes still have a negative binomial distribution, but with parameters α and $1 + n\beta$. The parameter estimation method was also modified. Instead of using method of moments, the following two equations are used to estimate the model parameters.

$$c_1/n = (1 + n\beta)^{-(1+\alpha)} \tag{1}$$

Table 2. Averages and standard deviations of $\hat{U}_p$ by P-G model based on 1,000 samples

	Example 1		Example 2		Example 3	
	$U_p = 222$		$U_p = 3{,}105$		$U_p = 278$	
f	avg	sd	avg	sd	avg	sd
0.01	647	97.96	12,399	706.02	1,440	336.17
0.05	519	26.74	4,654	135.65	447	50.87
0.1	415	13.58	2,557	59.26	285	21.85
0.5	101	1.20	404	4.39	78	2.52
1.0	39	0.0	157	0.0	41	0.0

and

$$K\alpha\beta = 1 \tag{2}$$

(See Appendix A.2 of Skinner et al. (1994) for a more detailed discussion.) Unfortunately, the modification did not improve the performance of the P-G model significantly. The empirical results show that this procedure severely underestimates the number of population uniques as the sample unique proportion increases beyond 15 per cent. They concluded that it is necessary to question the P-G model and its assumptions.

The modified P-G model has another problem. The root of Equation 1 (after substituting β using Equation 2) does not always exist. For a given set of K and n, the right side of Equation 1 has a maximum value which may be smaller than the observed unique proportions in samples. For instance, with $K = 1,024$ and $f = 0.05$ (i.e., $n = 4{,}398$) in Example 1 data set, the maximum value of the right side of the equation is 0.03577. But the maximum observed proportions in 1,000 samples is 0.04502.

After careful study of the P-G model, two problems become evident. First, the Poisson-Gamma model does not adequately represent the cell frequency structure in the population. This was indicated by the fact that the model greatly underestimates the number of population uniques with a 100 per cent sampling fraction. One reason for this problem is that $P(Y_i = 0)$ may be quite large even though Y_i are constrained to be strictly positive in the definition of K. For a Poisson distribution with parameter λ, the upper limit of $P(Y_i = 1)$ is obtained when $\lambda = 1$. Therefore, the maximum value of $P(Y_1 = 1)$ is $e^{-1} = 0.3679$, yielding a maximum possible value for U_p of $0.3679K$. In practice, we have found that the number of population uniques can greatly exceed $0.3679K$. For instance, in Example 2 the number of uniques is $0.4724K$. Greenberg and Zayatz (1992) have found the number of population uniques to be as high as $0.778K$. One approach to improving the model would be to allow K to include zero population counts in cells which are at least possible, in principle, and to fit a truncated version of the model to the positive sample cell counts. This and an alternative approach are discussed further in Section 4.2.

The second problem with this Poisson-Gamma model seems to be the method used to estimate the model parameters. It does not seem appropriate to use an estimation procedure based directly on the moments of sample cell frequencies, or simply replace N by n. When a random sample is drawn from the population, individual entities are drawn at random, not individual cells. An approach to this problem is discussed in Section 5.1.

4. Modeling the Cell Frequency Distribution

4.1. Overall population model

The search for the new estimation method started with the formulation of a model to represent the cell frequency structure in the population. To model this structure, the population is treated as a realization of a super population with K_{sp} cells.

The probability of the ith cell in the superpopulation will be denoted as $\pi_i, i = 1, \ldots, K_{sp}$ with

$$0 < \pi_i < 1$$

If every key variable had a uniform distribution, then the probabilities, π_i's, for all the cells should be the same. This is rarely the case in practice. For relatively large K_{sp}, the π_i's can be thought of as being spread between 0 and 1 continuously. Therefore, $\pi_i, i = 1, 2, \ldots, K_{sp}$ can be treated as continuous and identically distributed random variables with probability density function $f_\Pi(\pi)$.

The population of size N drawn from a superpopulation can be viewed as an outcome of N independent trials. In each trial, entities from each cell in the superpopulation get into the population independently according to the corresponding cell probabilities. So, the number of entities of the ith cell in the population, Y_i, will approximately follow a binomial distribution with parameters π_i and $N, i = 1, 2, \ldots, K_{sp}$.

This set-up is similar to the set-up of the model for word usage frequency problems (Sichel 1975; Bunge and Fitzpatrick 1993). The population cell frequencies, Y_i, are identically distributed random variables with probability distribution

$$P(Y = y|N) = \int_0^1 \binom{N}{y} \pi^y (1-\pi)^{N-y} f_\Pi(\pi) d\pi$$

Since the population size N is usually quite large and the π is small (due to the large K_{sp}), a Poisson distribution can be used to approximate the binomial distribution with $\lambda = N\pi$. Then

$$\begin{aligned} P(Y = y|N) &\approx \int_0^1 \frac{1}{y!} e^{-N\pi} (N\pi)^y f_\Pi(\pi) d\pi \\ &= \int_0^N \frac{1}{y!} e^{-\lambda} \lambda^y f_\Lambda(\lambda) d\lambda \end{aligned}$$

This results in a compound Poisson distribution as a model for the population cell frequencies. The only population cells from which sample units can be drawn are those where $Y_i > 0$. Therefore, the population is defined to have $K \le K_{sp}$ cells such that $Y_i \ge 1, i = 1, 2, \ldots, K$. The model developed here focuses on modeling these K non-zero population cell frequencies.

The distribution of λ needs to be chosen. It should be selected to describe the variation of the expected cell frequencies or sizes. We have found that, in practice, the frequency distribution of population cell size tends to have an inverse-J shape with a heavy upper tail. Figure 1 displays the cell size distributions for the population data sets of the five examples. To illustrate the inverse-J portion of the distributions in detail, the distributions of the first 30 cell sizes of Examples 1, 2, and 3 are also shown. Figure 1 shows that the shapes of the cell size distributions of the first four examples are very similar. As mentioned in Section 2, the cell size distribution of Example 5, which was used to illustrate an extreme situation, does not have the heavy upper tail.

To model the entire distribution, a mixture distribution for the λ is probably necessary. One distribution could be used to model the inverse-J shape while another distribution would model the long tail. Since the goal of this research is to estimate the number of unique cells in a population, finding a distribution that adequately models the inverse-J portion of the cell frequency, i.e., the frequency distribution of small cell sizes, may be more important than finding an appropriate distribution that would also model the tail.

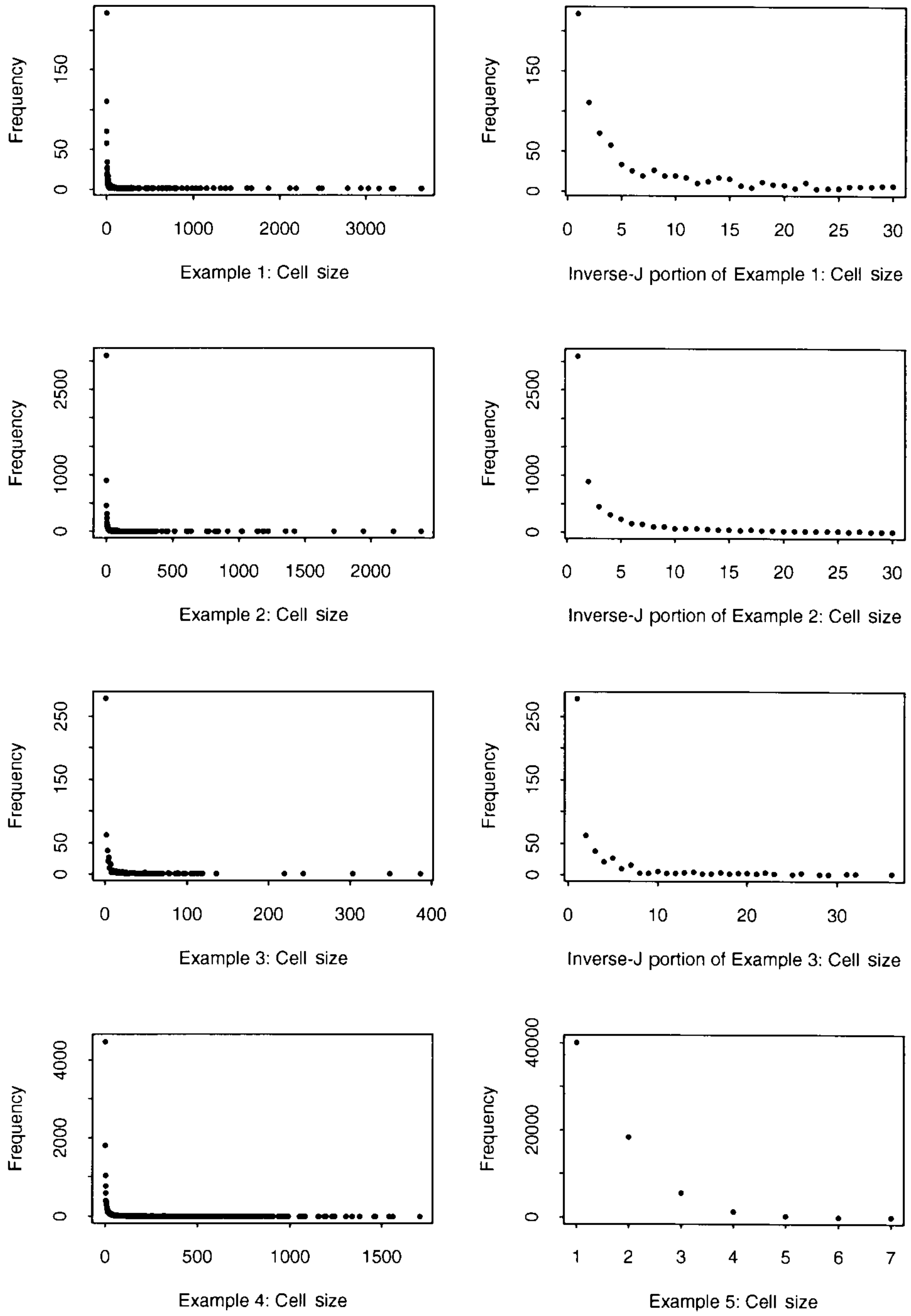

Fig. 1. Population cell size distributions. The first three pairs of the graphs display the cell size distributions and the details of the inverse-J portion of the distributions for Examples 1, 2, and 3. The graphs in the fourth row display the cell size distributions of Examples 4 and 5. Note the short tail of the distribution for Example 5

This will be the case provided the method of parameter estimation only uses information from the inverse-J portion of the entire distribution.

In the model developed here, it is assumed that λ has a single distribution for the entire range of the cell sizes. This distribution is chosen to accurately represent the frequencies of small cell sizes, i.e., inverse-J shape part, but not necessarily represent the long tail. A parameter estimation method is then developed to minimize the effect of the lack of fit for the tail distribution.

A Gamma distribution with parameters α and β is selected for the distribution of λ. The marginal distribution of the cell frequency in the population is then a negative binomial distribution. One advantage to this distribution is that the density function has a workable, closed form. With some modification, the compound Poisson-Gamma distribution seems to fit most real and simulated data sets we have worked with.

4.2. Modeling the inverse-J curve

The possible values of the cell frequencies for the K cells in the population are positive values, therefore the model distribution must exclude the possibility of $Y_i = 0$. There are two ways to exclude zero from the compound Poisson-Gamma distribution. One way is to truncate the distribution at zero. The second way is to slide the entire distribution one unit to the right.

Applying the zero-truncated negative binomial distributions to several population data sets indicated that this distribution did not adequately model the inverse-J part of the population distribution. In the inverse-J part of the distribution, the number of cells seems to decrease exponentially as the cell size increases. The rate of the decrease between the frequencies of size one and size two cells is quite high, and is much greater than the rate of decrease between the frequencies of size two and size three cells. The zero-truncated negative binomial distribution does not seem to follow these rates of decrease. The shape of the standard negative binomial distribution seems to follow the rate of the decrease between its first two points, i.e., zero and one. Truncating at zero causes a probability mass redistribution that de-emphasizes this drop.

A way to maintain the desired drop and still exclude zero is to slide the entire distribution one unit to the right. With this shift, the variable values start from one with the same probability density as the variable value zero in the standard negative binomial distribution. This distribution will be defined as the Slide Negative Binomial (SNB) distribution. The pdf of SNB is

$$P(Y = y) = \frac{\Gamma(\alpha + (y - 1))}{\Gamma(\alpha)(y - 1)!}\beta^{\alpha}(1 - \beta)^{y-1}, \quad y = 1, 2, \ldots$$

Using the SNB distribution to exclude zero from the compound Poisson-Gamma distribution, the model is set up as follows. Let $Y_1, Y_2, \ldots, Y_K$ be identically distributed random variables which represent the cell frequencies of the cells in the population for which $Y_i > 0, (K < K_{sp})$. Assume $Y_i, i = 1, 2, \ldots, K$ has SNB distribution with parameters α and β. The expected number of uniques in the population is

$$E(U_p) = KP[Y_1 = 1] = K\beta^{\alpha}$$

This expectation can be estimated as

$$\hat{U}_p = K\hat{\beta}^{\hat{\alpha}}$$

where K is the total number of non-zero cells in the population, and $\hat{\alpha}$ and $\hat{\beta}$ are estimates of α and β from SNB, respectively. In the next section, we will develop a method of estimation for α and β.

5. Population Parameter Estimation

5.1. Sample cell frequency distribution

Recall that the population is a realization of the superpopulation and has been treated as a simple random sample generated through binomial sampling from the superpopulation. The population is said to be finite in the sense that there is not an infinite number of entities in each population cell. Therefore, the frequency distribution of a cell in the sample depends on the particular outcome of this cell from that random process, i.e., the size of this cell in the population. The size of a cell in the population provides an upper limit for the size of this cell in the sample. For example, if a cell in the population has size one, the only possible sizes for this cell in the sample is one or zero. The distribution of the cell frequency in a sample needs to be based on a conditional distribution given a particular outcome of the population.

Consider a cell of size y in the population and let f be the sampling fraction. Since each entity of the cell has probability f to get into the sample, the size distribution of this cell in the sample will approximately follow a binomial distribution with parameters f and y. Let X denote the size of a cell in the sample, then

$$P(X = x|y) = \binom{y}{x} f^x (1-f)^{y-x}, \quad x = 0, 1, \ldots, y$$

The marginal distribution of X is

$$\begin{aligned} P(X = x) &= \sum_{y=max(x,1)}^{\infty} P(X = x|y)P(Y = y) \\ &= \sum_{y=max(x,1)}^{\infty} \frac{y!}{x!(y-x)!} f^x (1-f)^{y-x} \frac{\Gamma(\alpha + (y-1))}{\Gamma(\alpha)(y-1)!} \beta^{1-\alpha}(1-\beta)^{y-1} \end{aligned}$$

After simplification,

$$P(X = x) = \begin{cases} \frac{\beta^{\alpha}(1-f)}{(1-(1-f)(1-\beta))^{\alpha}} & x = 0 \\ \frac{\beta^{\alpha} f^x}{\Gamma(\alpha) x!} \left[\frac{\Gamma(\alpha+x)}{(1-(1-f)(1-\beta))^{\alpha+x}} (1-f)(1-\beta)^x + \frac{\Gamma(\alpha+x-1)}{(1-(1-f)(1-\beta))^{\alpha+x-1}} x (1-\beta)^{x-1} \right] & x = 1, 2, \ldots \end{cases} \tag{3}$$

Based on the previous assumptions, Equation 3 is the distribution of the sample cell frequencies.

5.2. Parameter estimation procedure

Given the sampling distribution for the cell frequencies in Equation 3, existing parameter estimation techniques could be applied. First consider the standard method of moments. The expectations of cell sizes can be derived using the probability distribution in Equation 3. The mean and variance are

$$E(X) = f\left[\frac{\alpha}{\beta}(1-\beta)+1\right] \tag{4}$$

and

$$Var(X) = \frac{\alpha f^2(1-\beta)}{\beta^2} + \frac{\alpha f(1-f)(1-\beta)}{\beta} + f(1-f) \tag{5}$$

Replacing the left sides of Equations 4 and 5 by the sample mean and variance of cell sizes, α and β can be estimated by simultaneously solving these two non-linear equations. Note that when the sample mean and variance are computed, the zero-size cells in a sample should be counted.

This estimation method was applied to the example data sets. The results showed that U_p tended to be underestimated for small sampling fractions and overestimated as the sampling fraction increases. The estimator appears to be inconsistent. A possible reason for this problem is that the population model does not fit the data well at the upper tail and this estimation method relies on information from the entire distribution. Recall that a distribution was chosen to model the inverse-J portion of the population distribution well, but not necessarily the long tail. Therefore, a different estimation procedure which uses information primarily from the inverse-J part of the population distribution needs to be used.

In a sample, the expected number of size one cells is

$$E(C_1) = KP(X=1)$$

and the expected number of size two cells is

$$E(C_2) = KP(X=2)$$

The parameters α and β can be estimated by setting these expectations to their observed sample values c_1 and c_2 and simultaneously solving the nonlinear equations. These equations are

$$c_1 = Kf\left[\frac{\beta}{1-(1-f)(1-\beta)}\right]^{\alpha}\left[\frac{\alpha(1-f)(1-\beta)}{1-(1-f)(1-\beta)}+1\right] \tag{6}$$

and

$$c_2 = K\frac{\alpha\beta^{\alpha}f^2(1-\beta)}{2[1-(1-f)(1-\beta)]^{\alpha+2}}[2-(1-\alpha)(1-\beta)(1-f)] \tag{7}$$

There is no closed form solution for this method of moments estimation. Therefore, a numerical procedure must be used to find $\hat{\alpha}$ and $\hat{\beta}$.

A method of moments estimation procedure using $E(C_1)$ and $E(C_2)$ is not the only choice in this problem. Any set of two simultaneous equations involving α and β could be used. There are several reasons why $E(C_1)$, and $E(C_2)$ were chosen. The lack of fit on the upper tail of the model will have less effect on the c_1 and c_2 values because small cells in the sample will be principally generated from small cells in the population. The inverse-J shape of the population is dominated by the probabilities associated with size one and size two cells. Therefore c_1 and c_2 contain much more information about this part of the population distribution than any other cell sizes. Also, c_1 is the only observed value which directly contains the information about population uniques. Finally, these estimation procedures are consistent, the proof of which is straightforward.

The results from applying this estimation technique to 1,000 random samples from Examples 1 to 5 are given in Table 3. These empirical tests provide evidence that this estimation method works well for both small and large sampling fractions. Figure 2 provides a graphical comparison of the performance of this new procedure to the existing methods for Example 1, 2, and 3. The improvement is rather dramatic, although the estimates of U_p still have a slight upward bias.

Notice that the results for Example 5 appear unbiased. Recall that the cell frequency distribution of this data set does not have a long tail. The SNB model has an excellent fit to the entire cell frequency distribution in Example 5. The good performance of this example indicates that the lack of fit of the SNB model at the upper tail portion of the population cell frequency distribution may still be influencing the estimation of U_p, when the sampling fraction is small. Another possibility for the estimation bias could be because the procedure does not take into account the constraints $\Sigma_{i=1}^{K_{sp}} \pi_i = 1$, $\Sigma_{i=1}^{K} Y_i = N$, and $\Sigma_{i=1}^{K} X_i = n$.

In practice, the total number of cells in a population may not be known. Based on the sample cell frequency distribution derived in this section, K can be estimated simultaneously with U_p. The expected number of size zero cells in a sample, which is the expected difference between observed number of cells in the sample and K, may be used as the third equation for this purpose. Therefore, K, α, and β can be estimated simultaneously by solving the three non-linear equations. The empirical results of this procedure indicate that this estimation method performs reasonably well for large sampling fractions. But the results are not stable for small sampling fractions. For small sampling fractions, the variance of the estimates of K can be quite large causing a large variation in the estimates of U_p.

6. Variance of $\hat{U}_p$

The estimator of U_p is a function of $\hat{\alpha}$ and $\hat{\beta}$, which in turn are functions of C_1 and C_2.

Table 3. Averages and standard deviations of $\hat{U}_p$ by SNB model based on 1,000 samples

	Example 1		Example 2		Example 3		Example 4		Example 5	
	$U_p = 222$		$U_p = 3{,}105$		$U_p = 278$		$U_p = 4{,}479$		$U_p = 40{,}217$	
f	avg	sd	avg	sd	avg	sd	avg	sd	avg	sd
0.01	382	68.52	4,055	301.46	321	79.92	6,882	291.0	45,055	2,957.1
0.05	356	53.34	3,648	161.79	305	45.91	6,524	164.2	40,765	1,439.6
0.1	347	63.76	3,645	105.68	307	39.06	6,395	242.4	40,374	1,140.9
0.5	238	28.51	3,200	90.12	280	25.62	4,691	121.9	40,143	165.3
1.0	222	0.0	3,105	0.0	278	0.0	4,479	0.0	40,217	0.0

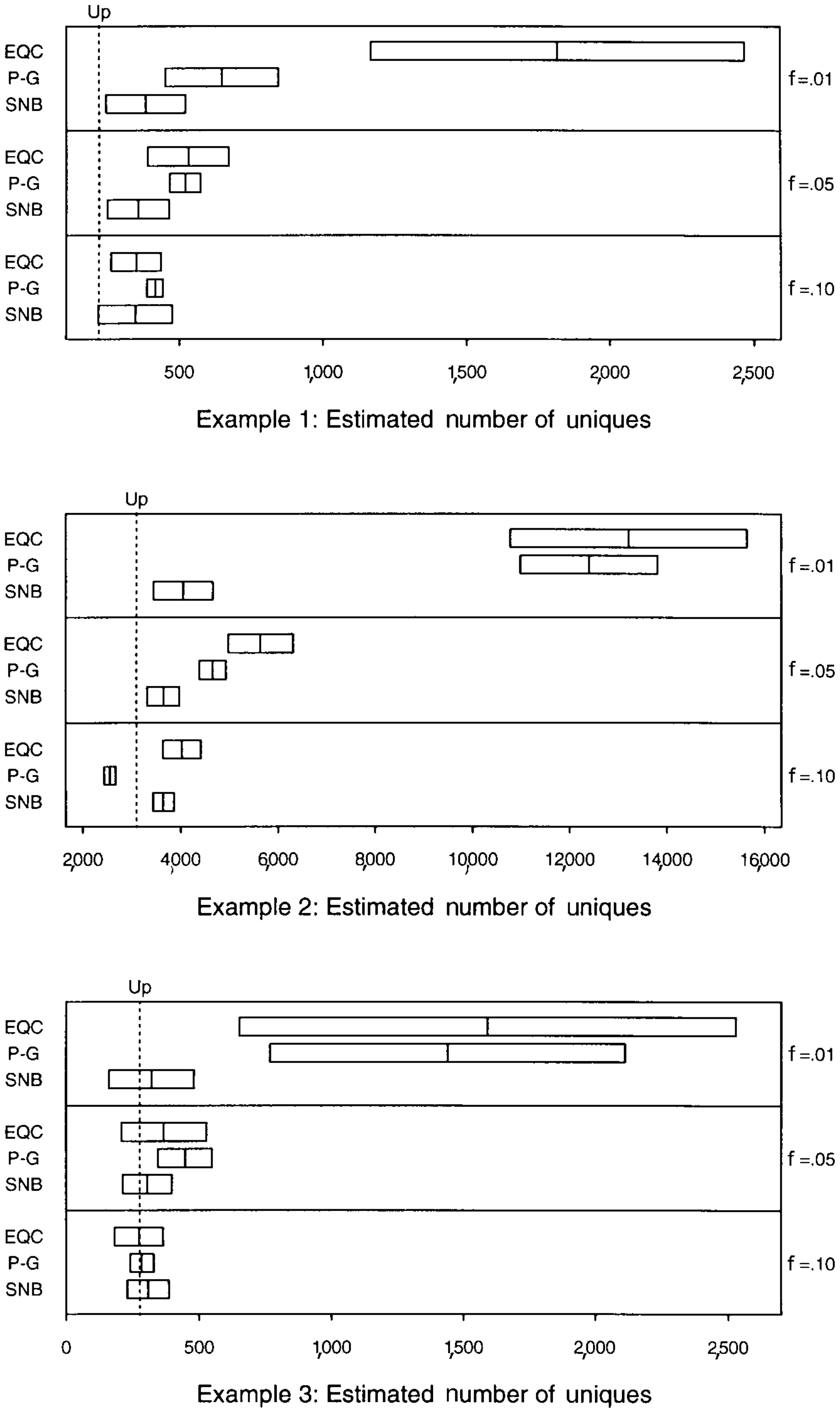

Fig. 2. Estimation method comparison. The middle lines of each box represent the averages for $\hat{U}_p$ of the 1,000 samples from each of the three sampling fractions. The two end lines of each box represent average ± 2 (sample standard deviation)

According to the δ-method, the variance of $\hat{U}_p$ can be estimated by

$$Var(\hat{U}_p) = [\partial\hat{U}_p]'V(\mathbf{C})[\partial\hat{U}_p]$$

where $[\partial\hat{U}_p]' = [\partial\hat{U}_p/\partial C_1, \partial\hat{U}_p/\partial C_2]$, and V(**C**) is the covariance matrix of $(C_1, C_2)'$. Recall that, during sampling, each cell may become a size one cell in the sample with probability $p_1 = P(X = 1)$ and may become a size two cell in the sample with probability $p_2 = P(X = 2)$. Therefore, the total number of sample size one cells, C_1, and the total number of sample size two cells, C_2, follow binomial distributions with parameters (K, p_1) and (K, p_2), respectively. The covariance matrix of **C** is

$$V(\mathbf{C}) = \begin{pmatrix} Kp_1(1-p_1) & Kp_1p_2 \\ Kp_1p_2 & Kp_2(1-p_2) \end{pmatrix}$$

The partial derivatives, $\partial\hat{U}_p/\partial C_1$ and $\partial\hat{U}_p/\partial C_2$, need to be calculated, via $\hat{\alpha}$ and $\hat{\beta}$, based on Equations 6 and 7. Because $\hat{\alpha}$ and $\hat{\beta}$ cannot be written explicitly as functions of C_1 and C_2, the rules for implicit differentiation must be applied. These details are given in Appendix A.

This approach was applied to Examples 1 to 4. A standard error estimate of $\hat{U}_p$ was calculated for each of the 1,000 samples from each sampling fraction. The means of the sets of 1,000 standard error estimates are listed in Table 4 in the column under the headings ''δ-meth.'' The empirical standard deviations given in Table 3 are listed again in this table in the columns under the headings ''sample.'' The results show that the δ-method is not sensitive to the sampling fraction. This insensitivity results in a significant underestimation when f is 0.01, and a slight overestimation when f is large (e.g., 0.5). This method was also applied to a set of samples of size 10,000 from Example 2 data. A similar pattern was observed.

7. Conclusions

The Slide Negative Binomial model developed in this work has significantly improved the results of population uniqueness estimation, especially for small sampling fractions. However, there is still an overestimation problem for small sampling fractions. This overestimation can lead to data being unduly withheld from legitimate users. For future study, a distribution or a mixture of distributions which can better model the upper tail of the population cell frequency distributions should be considered. The lack of independence

Table 4. Estimated standard errors of $\hat{U}_p$ using δ-method

	Example 1		Example 2		Example 3		Example 4	
f	sample	δ-meth	sample	δ-meth	sample	δ-meth	sample	δ-meth
0.01	68.52	48.25	301.46	157.7	79.92	46.94	291.0	182.3
0.05	53.34	52.36	161.79	123.3	45.91	45.92	164.2	185.3
0.1	63.76	65.19	105.68	120.3	39.06	41.99	242.4	221.3
0.5	28.51	39.14	90.12	109.3	25.62	33.21	121.9	161.2

in the Y_i's that leads to the constraints on the $\Sigma\pi_i$, ΣY_i, and Σx_i should also be incorporated into the model.

Another problem which needs further research efforts in this area is the estimation for the K unknown case. An investigation of the relationship between the values of K and the estimates of U_p may be helpful for determining the upper and lower bounds of $\hat{U}_p$ since, in practice, users might have some idea about the range of K even when the exact value of K is not known.

In addition to the estimation of population uniques, the SNB model may have more applications. It could be used to model the complete frequency distribution of a cross tabulation table provided an improvement to the fit at the upper tail of the distribution can be found. It may also be applicable to the problem of domain reduction dependencies in database systems (Hansen and Unger 1991).

Appendix A. Calculation Details for Estimating Variance of $\hat{U}_p$

To estimate the variance of $\hat{U}_p$ using δ-method, the partial derivatives $\partial\hat{U}_p/\partial C_1$ and $\partial\hat{U}_p/\partial C_2$ need to be calculated. Because U_p is estimated through the estimates of model parameters, α and β, these derivatives should be calculated as follows.

$$\frac{\partial\hat{U}_p}{\partial C_i} = \frac{\partial\hat{U}_p}{\partial\alpha}\frac{\partial\alpha}{\partial C_i} + \frac{\partial\hat{U}_p}{\partial\beta}\frac{\partial\beta}{\partial C_i}, \quad i = 1,2 \tag{8}$$

From $E(U_p)$ in Section 4.2., we have

$$\frac{\partial\hat{U}_p}{\partial\alpha} = K\beta^{\alpha}\ln\beta \quad \text{and} \quad \frac{\partial\hat{U}_p}{\partial\beta} = K\alpha\beta^{(\alpha-1)}$$

However, the estimates of α and β cannot be written explicitly as functions of C_1 and C_2. The derivatives $\partial\alpha/\partial C_i$ and $\partial\beta/\partial C_i$, $i = 1,2$, therefore, need to be calculated based on Equations 6 and 7 using the rules for implicit differentiation. Let $g_1(\alpha,\beta)$ and $g_2(\alpha,\beta)$ denote the right hand sides of Equations 6 and 7, respectively. Define

$$f_1(C_1,\alpha,\beta) = \ln C_1 - \ln(g_1(\alpha,\beta))$$

$$f_2(C_2,\alpha,\beta) = \ln C_2 - \ln(g_2(\alpha,\beta))$$

The reason for taking the natural logarithm is to simplify the calculation. Then, by taking partial derivative with respect to C_1 and C_2 on both sides of these two equations, four equations are formed. They are

$$\frac{\partial f_i}{\partial C_j} + \frac{\partial f_i}{\partial\alpha}\frac{\partial\alpha}{\partial C_j} + \frac{\partial f_i}{\partial\beta}\frac{\partial\beta}{\partial C_j} = 0, \quad \text{for} \quad i = 1,2; \quad j = 1,2$$

The four partial derivatives desired can be obtained through solving the equation set.

8. References

Adam, N.R. and Wortmann, J.C. (1989). Security-Control Methods for Statistical Database: A Comparative Study. ACM Computing Surveys, 21, 515–556.

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

Blien, U., Wirth, H., and Müller, M. (1992). Disclosure Risk for Microdata Stemming from Official Statistics. Statistica Neerlandica, 46, 69–82.

Bunge, J. and Fitzpatrick, M. (1993). Estimating the Number of Species: A Review. Journal of the American Statistical Association, 88, 364–373.

Denning, D.E. and Denning, P.J. (1979). Data Security. ACM Computing Surveys, 11, 227–249.

Denning, D.E., Denning, P.J., and Schwartz, M.D. (1979). The Tracker: A Threat to Statistical Database Security. ACM Transactions on Database Systems, 4, 76–96.

Duncan, G. and Lambert, D. (1986). Disclosure Limited Data Dissemination. Journal of the American Statistical Association, 81, 10–18.

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

Duncan, G. and Pearson, R. (1991). Enhancing Access to Microdata While Protecting Confidentiality: Prospects for the Future. Statistical Science, 6, 217–239.

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

Greenberg, B.G. and Zayatz, L.V. (1992). Measuring Risk in Public Use Microdata Files. Statistica Neerlandica, 46, 33–48.

Hansen, S.C. and Unger, E.A. (1991). An Extended Memoryless Inference Control Model: Accounting for Dependencies in Table Level Controls. Proceedings of ACM Sig-Management of Data Conference, May.

Keller, W.J. and Bethlehem, J.G. (1992). Disclosure Protection of Microdata: Problems and Solutions. Statistica Neerlandica, 46, 5–19.

Keller-McNulty, S. and Unger, E.A. (1993). Database System: Inferential Security. Journal of Official Statistics, 9, 475–500.

Kendall, M.S. (1980). Multivariate Analysis. Charles Griffin and Company, 34.

Lambert, D. (1993). Measures of Disclosure Risk and Harm. Journal of Official Statistics, 9, 313–331.

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

Sichel, H.S. (1975). On a Distribution Law for Word Frequencies. Journal of the American Statistical Association, 70, 542–547.

Skinner, C.J. (1992). On Identification Disclosure and Prediction Disclosure for Microdata. Statistica Neerlandica, 46, 21–32.

Skinner, C.J. and Holmes, D.J. (1992). Modelling Population Uniqueness. Proceedings of the International Seminar on Statistical Confidentiality. Statistical Office of the European Communities, Luxembourg, 175–199.

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

Spruill, N.L. (1983). The Confidentiality and Analytic Usefulness of Masked Business Microdata. Proceedings of the American Statistical Association, Section on Survey Research Methods, 602–613.

Zayatz, L.V. (1991a). Estimation of the Per Cent of Unique Population Elements in a Microdata File Using the Sample. Statistical Research Division Report Serie, Census/SRD/RR-91/08.

Zayatz, L.V. (1991b). Estimation of the Number of Unique Population Elements Using a Sample. Proceedings of the American Statistical Association, Section on Survey Research Methods, 369–373.

Received October 1995
Revised June 1997