A Per-Record Risk of Disclosure Using a Poisson-Inverse Gaussian Regression Model

Michael Carlson*

December 18, 2002

Abstract

Per-record measures of disclosure risk have potential uses in statistical disclosure control programs as a means of identifying sensitive or atypical records in public-use microdata files. A measure intended for sample data based on the Poisson-inverse Gaussian distribution and overdispersed log-linear modeling is presented. An empirical example indicates that the proposed model performs approximately as well as the Poisson-lognormal model of Skinner and Holmes (1998) and may be a tractable alternative as the required computational effort is significantly smaller. It is also demonstrated how to extend the application to take into account population level information. The empirical results indicate that using population level information sharpens the risk measure.

Keywords: Disclosure control; Log-linear models; Poisson-inverse Gaussian; Risk-per-record; Uniqueness.

1 Introduction

A growing amount of literature dealing with various aspects of statistical disclosure control (SDC) has been published in the last decades of which we may refer the interested reader to e.g. Dalenius (1977), Duncan and Lambert (1989), Duncan and Pearson (1991), Fienberg (1994), Frank (1976, 1988), Lambert (1993), Skinner et al. (1994), Willenborg and de Waal, (1996, 2000). Recent publications include Doyle et al. (2001) and Domingo-Ferrer (2002).

*Department of Statistics, Stockholm University, SE-106 91 Stockholm, Sweden. E-mail: Michael.Carlson@stat.su.se

A special case concerns public-use microdata files which are often released to researchers so that they may conduct their own analysis. A microdata file is defined as consisting of records corresponding to individual units in a sample from a finite population. If the data originates from e.g. a census or large database, only a subset of the complete set is sometimes considered for release so we will only consider cases where a sample is available. Each record consists of a number of attributes pertaining to the individual scores of the corresponding unit. Clearly there are potential risks involved concerning the possibility of disclosing individual information about the respondents from which the data is derived. Even when the file has been anonymized by the removal of direct identifiers such as names and personal ID-numbers etc., the risk remains since a perceived intruder may use a set of matchable attributes, or *key* attributes, included in the data to establish a link between a given record and some individual in the population.

The disclosure risk may be conceived of as the evidence in support of a correct link between a record in the file and a unit in the population. Methods that assign each record an individual measure of disclosure risk, based on the characteristics of the record, have a number of potential uses. As with any SDC technique it is desirable that the amount of alterations made to the original data is small enough to ensure that the effect on subsequent analyses is within reasonable limits. The agency releasing the data could for example focus on the highest ranking records and apply disclosure control methods to these separately, e.g. by local suppression, rather than modifying the whole data set, e.g. by global recoding. Such a practice would be analogous to, and easily incorporated with, other data processing procedures, such as editing. See also the discussion in Willenborg and de Waal (1996 p. 137).

The framework considered in the present paper is basically identical to the method proposed in Skinner and Holmes (1998) who based their risk-per-record measure on the concept of uniqueness. A unique is simply defined as an entity with a unique combination of scores on the set of key attributes. A *population unique* is unique in the population and a *sample unique* is accordingly unique in the sample. The idea is to assess the probability that a unique record is also unique in the population. Skinner and Holmes do this by taking into account the individual scores of the key attributes for each record using simple overdispersed log-linear models. Skinner and Holmes used the Poisson-lognormal distribution (PLN) as the basis for their model, here we investigate the performance of the Poisson-inverse Gaussian (PiG).

The PiG introduced by Holla (1966) in studies of repeated accidents and recurrent disease symptoms and has since been applied to sentence-length and word frequency data and model repeat-buying behavior, (Sichel, 1974, 1975, 1982), species abundance (Ord and Whitmore, 1986) and insurance

claim data (Willmot, 1987). It has been noted that the frequency distribution in disclosure applications been tends to have an inverse J-shape with heavy upper tail (cf. Chen and Keller-McNulty, 1998). Since the PiG distribution is characterized by its positive skewness and heavy upper tail it appears to be an appropriate distribution for modeling frequency counts in disclosure applications. It was applied to the disclosure problem in Carlson (2002) were it was shown to provide good estimates in an empirical example. A disadvantage with the lognormal is that it is not available in closed form and that numeric integration is required whereas the PiG is easily defined by simple recursion formulas and the derived risk measure is in closed form. The PLN on the other hand is supported by reasonable theoretical justifications and strong empirical evidence, see Skinner and Holmes (1993, 1998) and Marsh et al. (1994).

In the Nordic and many other countries detailed population statistics are frequently being published from registers and population uniques can either be inferred or excluded directly from the published tables or the published tables can be used as auxiliary information along with the sample data. Marginal distributions of potential key variables are often readily available from public sources and even two-way tables would not necessarily be considered hard to come by if the variables are not to exotic. An extension of the method which more or less suggests itself is to use population level margins or two-way tables in the estimation of the model parameters.

Thus, the scope of the study is two-fold: to empirically evaluate the PiG model as a simpler alternative to the PLN and to investigate if and to what extent auxiliary information in the form of known population margins or two-way tables improve the risk-per-record measure. The paper is organized as follows. In the following section we introduce some basic notation and in section 3 we briefly discuss re-identification risks. In section 4 the basic PiG model is developed. In section 5 we apply the method to microdata samples using the sample information with and without population level information. Some concluding remarks are given in section 6.

2 Basic Setup and Notation

Consider a finite population U of size N from which a simple random sample $s \subseteq U$ of size $n \leqslant N$ is drawn. The sampling fraction is denoted by $\pi_s = n/N$. The sample units each correspond to a record in the microdata file considered for release. With each unit in the population is associated the values of a set of discrete *key variables*, $X_1, \ldots, X_q$ with $C_1, \ldots, C_q$ categories, or levels, respectively. The *key* is defined by the cross-classification of the key variables

and is denoted by X. There are $\Pi C_i = C$ possible combinations of values of the key variables and for simplicity we let the different combinations, or cells, be labeled by x, i.e. we let $x = (x_1, x_2, \ldots, x_q)$, where $1 \leqslant x_j \leqslant C_j$ for $j = 1, \ldots, q$. Thus, the key partitions the population into C subpopulations $U_x \subseteq U$ and by F_x we denote the number of units belonging to subpopulation U_x, i.e. the population frequency or size of cell x for $x = 1, \ldots, C$. The sample counterpart is denoted by f_x and it is clear that

$$\sum_{i=1}^{C} F_x = N, \qquad \sum_{i=1}^{C} f_x = n.$$

Of these quantities, C, N and n are usually fixed by design, the f_x are observed and the F_x are assumed to be unknown. The aim is to model and estimate the population frequency structure, with special attention to those F_x which are of size one. The number of cells in the population with only one unit, will be denoted by T_1 and constitutes the number of population uniques. The sample counterpart, is denoted by t_1 which is the number of sample uniques. Also, it is convenient to denote the number of population uniques falling into the sample by $t_{1,1}$ and we note that $t_{1,1} \leqslant t_1$.

In disclosure applications the number of combinations of key variable scores C is usually quite large, and it is almost inevitable that a large number of cells will not be observed simply by chance. It is also obvious in many situations that certain combinations of the key variables may be impossible, such as married 4-year olds or male primiparas, i.e. so-called structural zeroes. Since the number of structural zeroes often is assumed unknown, special considerations, for instance in estimation, may be called for. See e.g. Skinner and Holmes (1993).

3 Re-identification Risk

The basic definition of the disclosure problem considered here is the same as that of many other authors, e.g. Bethlehem et al. (1990), Elliot et al. (1998), and Skinner and Holmes (1998). Consider an intruder who attempts to disclose information about a set of identifiable units in the population termed *targets*. The intruder is assumed to have prior information about the key values of the targets and attempts to establish a link between these and individual records in the released microdata file using the values of the key attributes. We assume further that there is no measurement error, which could lead to false matches, that the key variables are discrete and that units are included in the file with equal probability.

Now assume that the intruder finds that a specific record r in the microdata file matches a target with respect to the key X. Now F_x is the number of units belonging to subpopulation U_x and we let $x(r)$ denote the value of X for record r. If $F_{x(r)}$ was known the intruder could infer that the probability of a correct link is $F_{x(r)}^{-1}$ and if $F_{x(r)} = 1$ the link is correct with absolute certainty under the assumption of no measurement error. Usually the intruder will not know the true value of $F_{x(r)}$ since the microdata set contains only a sample but by introducing a superpopulation model he may attach a probability distribution $Pr(F_x = j)$ to the cell frequencies.

Furthermore, it could be argued that an intruder will be more inclined to focus on records that are sample unique since it is only these that can by definition be population uniques. An indicator of the identification risk is obviously the number or proportion of population uniques included in the sample amongst the sample uniques. This type of measure is considered by e.g. Elliot et al. (1998), Fienberg and Makov (1998), and Skinner and Holmes (1998). A variety of methods based on superpopulation models, especially compound Poisson models, have been proposed in the literature. Bethlehem et al. (1990) were perhaps the first to adapt the superpopulation approach and have since been followed by among others Skinner and Holmes (1993), Chen and Keller-McNulty (1998), Samuels (1998), Hoshino (2001), and Carlson (2002). Under a superpopulation model it is assumed that the population at hand, as defined by the frequency structure of the key attributes, has been generated by some appropriate distribution. The risk assessment, here in terms of uniqueness, is then reduced to a matter of parameter estimation and prediction.

Now assume for example that the file holds two 42-year old female physicians amongst the records, one living in a large urban area, the other in a small rural community (age, sex, occupation and geographical area are assumed to be attributes included in the data). Assume further that both are sample uniques. Because of the intuitive "rareness" of the latter it seems natural to assign a larger risk to this record than to former. The proportion of population uniques amongst sample uniques is however constant for all sample uniques and works only as an indicator of disclosure risk for the file as a whole or as a *file-level* risk measure. To define a *record-level* measure of risk it seems desirable to condition on the values of the key variables defining the key. Furthermore, from the intruders viewpoint it should be considered optimal to utilize as much information as possible, such as the structure of the data inherent from the variables defining the key. Ideally such a measure should be close to one for records that are population unique and close to zero for all other records. Thus, under some suitable model, it follows naturally to define the individual risk of a sample unique record as the conditional

probability given the key, i.e.

$$\text{Risk}\,(x) = \Pr\left(F_X = 1 \mid f_X = 1; X = x\right). \tag{1}$$

Some propositions in this direction have been published in recent years. In Elliot (2002) a non-parametric method based on a bootstrap argument is proposed. Fienberg and Makov (1998) and Skinner and Holmes (1998) both use standard log-linear models to evaluate per-record risks and Takemura (2001) considers fitting a Lancaster-type additive model of interaction terms. Benedetti et al. (1999) presents a method which results in taking into consideration instead the individual sampling weights and extends its application to hierarchical data. In this paper we adopt, as already mentioned, the approach of Skinner and Holmes.

4 Modeling the Cell Frequencies

4.1 Basic Model

As a starting point we assume that the cell frequencies are generated independently from Poisson distributions with individual rates λ_x, $x = 1, \ldots, C$. The Poisson model is motivated by thinking of the N units in the population as falling into the C different cells with probability of the ith cell denoted by π_x. Given the N, C and the π_x the frequencies will follow a multinomial distribution and if the number of cells is large enough the cell frequencies are approximately independent binomial with parameters N and π_x respectively. Since the population size is usually quite large and the π_x small due to large C the Poisson distribution is used to approximate the binomial with $\lambda_x = N\pi_x$.

A further assumption is to view the λ_x as independent realizations of independent continuous random variables Λ_x with probability density functions (pdf) $g_x(\lambda)$ which we define to depend on the given combination of key variable scores. The specification of the mixing distribution $g_x(\lambda)$ is the crucial step. Skinner and Holmes (1993, 1998) proposed the lognormal distribution and provided a theoretical justification for the model. In Carlson (2002) the Poisson-inverse Gaussian distribution was described as a possible alternative for file-level risk assessment and this model is extended in the following subsection to provide a record-level risk measure.

4.2 Poisson-inverse Gaussian Regression Model

With each cell count F_x we associate a set of covariates x and a Poisson regression model would stipulate that, given x, F_x is distributed as a Poisson

with mean μ_x. One way of introducing random effects or extra-Poisson variation into such a model is the common multiplicative Poisson random-effects model as described in Dean et al. (1989). The model is defined by

$$\Pr(F_x = j) = \int_0^\infty \frac{(\mu_x \nu)^j e^{-\mu_x \nu}}{j!} g(\nu)\, d\nu, \quad j = 0, 1, \ldots \tag{2}$$

where $g(\nu)$ is a probability density function and $\mu_x = \mu(x; \beta)$ is a positive-valued function of the covariates which depends on a vector β of unknown regression coefficients. Given the fixed values of the covariates x, and a random effect ν with density $g(\nu)$, $\nu > 0$, the cell frequency F_x has a Poisson distribution with mean $\mu_x \nu$. Furthermore, we may assume without loss of generality that $E(\nu) = 1$. This parametrization has the property that when μ_x takes the common log-linear form

$$\mu_x = \mu(x; \beta) = \exp(x'\beta) \tag{3}$$

random and fixed effects are added on the same exponential scale. For a general overview of log-linear models, see e.g. Christensen (1997). We will consider only two log-linear models in the examples of section 5: a simple *main effects only* model, (ME), and an *all two-way interactions* model, (TW).

Specifying the distribution of the random effect ν is the next step and using the parametrization of Dean et al. (1989) we consider the inverse-Gaussian (iG) distribution with density

$$g(\nu) = \frac{1}{\sqrt{2\pi\tau\nu^3}} \exp\left(-\frac{(\nu - 1)^2}{2\tau\nu}\right), \quad \nu > 0, \tag{4}$$

where $E(\nu) = 1$ and $Var(\nu) = \tau$, and where the parameter τ is assumed unknown. A review of the iG is given in Folks and Chhikara (1978). The distribution of F_x given x resulting from (2) is then a Poisson-inverse Gaussian (PiG) regression model, with mean μ_x and variance $\mu_x(1 + \mu_x\tau)$ respectively. A short review of the PiG model is given in Carlson (2002).

Given the fixed values of the covariates x, and a random effect ν with density (4) it follows that

$$F_x \mid x, \nu \sim Po(\nu \exp(x'\beta)).$$

Assuming Bernoulli sampling with sampling probability $\pi_s = n/N$ (cf. Särndal et al., 1992, chapter 3) it follows that

$$f_x \mid \mu_x \nu \sim Po(\pi_s \mu_x \nu) \quad \text{and} \quad F_x - f_x \mid \mu_x \nu \sim Po((1 - \pi_s)\mu_x \nu) \tag{5}$$

independently and that

$$f_x \mid F_x \sim Bin\left(F_x, \pi_s\right). \tag{6}$$

It is then easily seen that the marginal distribution of the sample cell frequencies f_x is also distributed as PiG with mean $\pi_s\mu_x$ and variance $\pi_s\mu_x\left(1+\pi_s\mu_x\tau\right)$. See section 2 of Sichel (1982) and the discussion concerning sampling in Takemura (1999). In the following $\pi_s\mu_x$ is denoted by μ_{xs}.

4.3 A Per-record Risk of Disclosure

It is now a simple matter to express the risk measure in (1) in terms of the model parameters. Given the present model the probabilities

$$\Pr\left(F_x=1\right)=\frac{\mu_x}{\sqrt{1+2\mu_x\tau}}\exp\left(\frac{1}{\tau}\left(1-\sqrt{1+2\mu_x\tau}\right)\right)$$

and

$$\Pr\left(f_x=1\right)=\frac{\mu_{xs}}{\sqrt{1+2\mu_{xs}\tau}}\exp\left(\frac{1}{\tau}\left(1-\sqrt{1+2\mu_{xs}\tau}\right)\right).$$

are easily derived (see Carlson, 2002, for details). Given (6) and the above, the individual risk-per-record measure is derived as

$$\begin{aligned}\text{Risk}\left(x\right) &= \Pr\left(F_x=1 \mid f_x=1;x\right)=\frac{\pi_s\Pr\left(F_x=1\right)}{\Pr\left(f_x=1\right)} \\ &= \frac{\sqrt{1+2\mu_{xs}\tau}}{\sqrt{1+2\mu_x\tau}}\exp\left(\frac{1}{\tau}\left(\sqrt{1+2\mu_{xs}\tau}-\sqrt{1+2\mu_x\tau}\right)\right) \end{aligned} \tag{7}$$

which is calculated for each unique record in the sample. The risk measure is estimated by replacing the parameters for their respective estimates. When τ, or σ^2 for the PLN model, is taken to be zero, (2) is reduced to a simpler Poisson regression model and a simplified risk measure is derived from (5) as

$$\Pr\left(F_x=1 \mid f_x=1;x\right)=\Pr\left(F_x-f_x=0;x\right)=\exp\left(-\left(1-\pi_s\right)\mu_x\right). \tag{8}$$

4.4 Estimation

Dean et al. (1989) derived the maximum likelihood and quasi-likelihood estimators for the PiG regression model in (2). However, the estimation procedure that we use in the examples of section 5 is identical to the ad hoc procedure described in Skinner and Holmes (1998). Rather than estimating

the regression coefficients in (3) and the PLN parameter σ^2 simultaneously using e.g. maximum likelihood, they first estimated the expected means μ_{xs} directly in the usual way for ordinary log-linear models. Once the μ_{xs} are estimated, σ^2 is estimated by a simple moment estimator. The procedure is an ad hoc approach as it does not fully account for the model specification in (2) but has the advantage of being easily implemented in standard software. As before we let x correspond to the values $x_1, \ldots, x_q$ of the respective key variables $X_1, \ldots, X_q$. Let $f_{x_j}^{(j)}$ denote the number of units in the sample taking the value x_j on the jth key variable X_j. For the main effects only log-linear model (ME), we then estimate the individual means μ_{xs} by n times the product of the marginal proportions of the respective key variable scores, i.e.

$$\hat{\mu}_{xs} = n \frac{f_{x_1}^{(1)}}{n} \cdots \frac{f_{x_q}^{(q)}}{n}.$$

For the all two-way interactions (TW) log-linear model, iterative proportionate fitting was used, see e.g. Christensen (1997, pp. 87-89). The procedure requires the counts of all observed two-way combinations of the key variables, i.e. we calculate $f_{x_j x_k}^{(j,k)}$ defined as the number of units in the sample taking the value x_j on the jth key variable X_j and value x_k on the kth key variable X_k for all possible combinations x_j, x_k. Note also that μ_x is estimated by $\hat{\mu}_{xs}/\pi_s$.

An obvious way to incorporate known population level information is to replace the corresponding sample counts for the population counts, e.g. for the ME-model one would use instead

$$\hat{\mu}_{xs} = n \frac{F_{x_1}^{(1)}}{N} \cdots \frac{F_{x_q}^{(q)}}{N}$$

where $F_{x_j}^{(j)}$ is the population level analogue to $f_{x_j}^{(j)}$. The modification for the TW-model is analogous by using the corresponding population two-way counts.

As mentioned in section 2, a problem may be that the number of structural zeros is usually unknown. To allow for this, a moment estimator of τ is defined by noting that the first two conditional moments of f_x are

$$E\left(f_x \mid f_x > 0\right) = \frac{\mu_{xs}}{1 - p_{0x}}$$

and

$$E\left(f_x^2 \mid f_x > 0\right) = \frac{\mu_{xs}\left(1 + \mu_{xs} + \mu_{xs}\tau\right)}{1 - p_{0x}}.$$

respectively and where p_{0x} denotes the probability under the model that cell x is empty in the sample. By combining these conditional moments it follows that

$$\frac{E\left(f_x^2 \mid f_x > 0\right) - E\left(f_x \mid f_x > 0\right)}{\mu_{xs} E\left(f_x \mid f_x > 0\right)} - 1 = \tau. \tag{9}$$

Hence, once the expected sample means μ_{xs} have been estimated, a simple estimator is obtained by substituting the μ_{xs} in (9) for their estimates yielding

$$\hat{\tau} = \frac{\sum_{x>0}\left(f_x^2 - f_x\right)}{\sum_{x>0} f_x \hat{\mu}_{xs}} - 1 \tag{10}$$

This is exactly the approach used by Skinner and Holmes (1998) for estimating the corresponding PLN parameter σ^2. In both cases the sums may be taken over all cells and not just those for which $f_x > 0$, since the values summed are both zero when $f_x = 0$. With this simplified approach it is possible to obtain negative estimates of τ (and σ^2) and if this occurs, τ may be taken to be zero. This is equivalent to reducing (2) to a simpler Poisson regression model, since then $\nu = 1$ with probability equal to one.

5 An Example

5.1 Description of the Data

A population consisting of individuals of ages 18 - 65 residing in three counties in the southern part of Sweden was compiled from the Store database, managed by the Swedish National Social Insurance Board (Riksförsäkringsverket, 2002). After removing individuals for which the marital status was unknown (code = 8), the total population size was $N = 268,607$. Six categorical variables were used: age in one year bands (48), sex (2), marital status (7), children , yes or no, (2), county (3), and income in 50,000 SEK bands and top-coded (20). The numbers in parenthesis indicate the number of observed categories of the respective variables from which the total number of possible combinations is given as $C = 80,640$. Of these, $T_0 = 69,185$ were found to be empty leaving a total of $11,455$ observed combinations. The number of population uniques T_1 was $2,607$ or approximately 0.97% of the population. The largest cell contained $1,426$ units (one cell).[1]

[1]It was later found that a few individuals in the data actually were deceased or had emigrated but for insurance reasons were still in the system. For our purposes this is however of little or no consequence as we are mainly illustrating the method.

From the data set two simple random samples without replacement were drawn. The first was of size $n = 5,373$ corresponding to $\pi_s = 0.02$. The largest observed cell size of the $2,612$ non-empty cells in the sample was 37 (one cell). The observed number of sample uniques was $t_1 = 1,560$ or approximately 29% of the sample. Of these, $t_{1,1} = 42$ were found to be population uniques or approximately 2.7% of the number of sample uniques. This is a bit short of the expected number of population uniques expected to fall in the sample, i.e. 2% of T_1. The second sample was of size $n = 26,861$ corresponding to $\pi_s = 0.10$. The largest of the $5,751$ non-empty cells in this sample was 154 (one cell) and the observed number of sample uniques was $t_1 = 2,304$ or approximately 8.6% of the sample. Of these, $t_{1,1} = 259$ where found to be population uniques which is a little closer to the expected number given the sampling fraction, i.e. approximately 10% of T_1. The population and sample frequency distributions of the cell sizes are given in table 1.

5.2 Using Only Sample Information

We first fitted the ME model to each of the two samples as described in the preceding section using only sample data. Two codes for marital status and six for income were not observed at all in the 2% sample. This meant that for this set, half of the possible combinations on the key variables were assumed to be structural zeroes, leaving a total of $40,320$ means $\hat{\mu}_{xs}$ to be estimated. For the 10% sample the same two codes for marital status and one for income were not observed in the sample, resulting in $54,720$ estimated means $\hat{\mu}_{xs}$ and the remaining $25,920$ combinations being taken as structural zeroes. The moment estimates of τ and σ^2 are given in tables 2 and 5. For each of the sample unique records, the risk measure (7) was then calculated.

We also calculated the PLN based risk measure of Skinner and Holmes (1998) for comparison. The PLN based risk measure requires numerical integration and we experimented with various variable substitutions of the lognormal kernel and different numeric integration techniques and settled for the transformation $\lambda = (1-t)/t$ to obtain finite integration limits and the Matlab (2001) `quadl` routine which uses an adaptive quadrature technique. The calculated probabilities were checked against and found to agree with the tabulated values in Grundy (1951). A problem with the PLN based measure for this data set was the occurrence of individual risk measures larger than one. This is due to the numerical precision of the integration procedure used. Upon inspection we found that these records all had the smallest estimated means $\hat{\mu}_{xs}$ amongst the sample uniques, approximately 2×10^{-4} for the 2% sample and 5×10^{-4} for the 10% sample. As it is clear that both the PiG and the PLN risk measures are monotonically decreasing functions of μ and

that the limit as μ approaches zero is one, we simply set the measures for these records to one.

Next we considered fitting the TW model. This involved using iterative proportionate fitting to fit the $\hat{\mu}_{xs}$ to agree with the $1,977$ possible two-way combinations. Of these, 943 in the 2% sample and 781 in the 10% sample turned out to have sample counts equal to zero. All combinations on the key variables corresponding to these zero counts were taken as structural zero cells which resulted in $13,890$ and $19,255$ estimated means $\hat{\mu}_{xs}$, respectively. With the 2% sample, the estimates of τ and σ^2 both turned out to be negative suggesting the simpler Poisson regression model and the simplified risk measure (8). Thus the PiG and PLN risk measures are equal for this sample. With the 10% sample, the estimate of τ was larger than zero whereas the estimate of σ^2 turned out to be negative, again suggesting the simplified risk measure (8) for the PLN model.

Once the respective models were fitted we calculated the individual PiG and PLN risk measures for each sample unique and these are plotted against each other in figure 1. When the ME model was used, we note that the PiG based measure is slightly larger than the PLN measure for PLN values below approximately 0.20-0.25 and slightly smaller for larger values. Generally, the PLN measures are more stretched out towards the endpoints of the range, whereas the PiG tends to gravitate slightly towards a point in the lower quarter of the range relative the PLN. The reason for this may be that the PLN measure is closer to the ideal measure (close to unity for population uniques and close to zero for all others) but could also be that the PLN overestimates the risks. Either way, the differences appear to small to draw any substantial conclusions from the present study. The similarity is however striking as the ranking order of the PiG risk measures is identical to the ranking order of the PLN based measures. This should not come as a surprise as the main factor determining the individual risks are the estimated means $\hat{\mu}_{xs}$ which are the same in both measures. Turning to the TW model there is no need for further comment with respect to the 2% sample as the measures are identical. As for the 10% sample the differences between the PiG and PLN are even smaller compared to the ME model. This is due to the estimate of τ being so close to zero, resulting in a PiG measure that nearly coincides with the simplified measure (8).

We also divided the individual risk measures into groups according to the ranges defined in tables 2 - 5 and recorded the number of sample uniques falling into each range and the respective percentages of these which were found to be either population uniques or one of a population pair, i.e. $F_x = 2$. As reported by Skinner and Holmes (1998) a relatively strong relationship between the risk measure and the proportion of population uniques within

each range is seen although the results are not as clear cut for either measure as in their study. This is the case for the smaller sample and the ME model where the percentages increase only approximately monotonically. The TW model, where the simplified measure was used, displays an even worse correspondence between the risk and the percentage of population uniques. We note that very few records have fallen into the higher ranges resulting in a large degree of instability. If for example only two observations fall into a given range, the only possible outcomes are 0%, 50% and 100% population uniques in this range. On the other hand this lack of relationship may from a disclosure point of view be considered favorable since the intruder will not find many high-risk records either, using the same framework. Turning to the larger sample we note that the correspondence is improved for both the ME and the TW models. As expected, more records have fallen into the higher ranges and this would account for an increased stability. Another problem with the examples considered here is the smaller sample sizes. In Skinner and Holmes' study a sample with approximately $45,000$ records was used, compared to the $5,373$ and $26,861$ in our examples. Especially for the smaller sample the amount of available information is not likely to be as sufficient when e.g. $1,034$ two-way proportions are being estimated.

In order to provide some more detail of the results, the PiG based risk measures are plotted against the true population cell size in figure (2). It is seen that the correspondence between the risk measure and the population cell size is quite strong. Records belonging to the largest population cell sizes are all among the lowest ranking risk measures whereas those belonging to smaller cell sizes are more evenly spread out across the entire range. The highest ranking risks are on the other hand all associated with relatively small cell sizes which is indicated also by the number of population pairs ($F_x = 2$) falling into the different ranges as seen in tables 2 - 5. The interpretation is that highest ranking records are most like likely either population uniques or one observation of a population pair.

5.3 Using Known Population Margins

We next considered using population level marginal counts in the estimation of the expected means μ_{xs}. There are of course many conceivable ways of combining information from both population and sample but we restricted the study to two settings: (1) all main population level margins are assumed known and the ME model is used, and (2) all population level two-way tables are assumed known and the TW model used. This was done with both the 2% and 10% samples.

In the first setting it was found that all levels of all the key variables were

observed which resulted in all of the $80,640$ possible key combinations being fitted. For the second setting is was found that 514 of the $1,977$ two-way margins were unobserved in the population. As expected from the sample data, these were mainly combinations involving marital status and higher levels of income. This left a total of $29,373$ combinations that were used in the iterative proportionate fitting procedure to fit the $\hat{\mu}_{xs}$. The estimates of τ from (10) were all positive save the 10% sample with the TW setting, where $\hat{\tau}$ was found to be negative and accordingly set to zero, suggesting the simplified measure (8).

The PiG based risk measure was calculated for each sample unique and the results of the study are given in tables 6 and 7. The main result is that the risk measure appears significantly improved. More records are falling into higher ranges, giving stability to the correspondence between the risk measure and the proportion of population uniques within each range. Furthermore, low-risk records have dropped to lower ranges which is seen e.g. in the increase of records in the lowest range. On the whole, the risk measure appears much sharper as compared with using only sample information.

6 Remarks

From the results of the study we conclude that the PiG based model was able to provide risk measures approximately equivalent to the PLN. However, more research is warranted, both theoretical and empirical, as the conclusion of equivalent measures is based on only one specific data set with one specified key. Larger differences may very well occur and it is vital to investigate the circumstances under which either of the two will work. On the other hand such differences can be accounted for by simply altering the threshold value above which records are flagged as sensitive as the ranking order is the same in both models.

It was noted that for smaller sample sizes both the PiG and PLN based measures appeared to be unstable, especially when the TW model was used. This shows that some further research with respect to model selection may be called for. Skinner and Holmes argue that to elaborate modeling may result in unstable estimates of the risks. There is also another consideration. When a model is formulated within the present framework, we actually do *not* want a perfect fit to the sample data since this would yield estimates equal to the observed cell frequencies, $\hat{\mu}_{xs} = f_x$ and especially for sample uniques, $\hat{\mu}_{xs} = 1$. It is easily seen that the estimate of τ (or σ^2 if the PLN is used) will always be negative, suggesting the simplified risk measure which

turns out to be constant over all sample uniques, i.e.

$$\Pr\left(F_x = 1 \mid f_x = 1; x\right) = \exp\left(-\left(1 - \pi_s\right)\hat{\mu}_{xs}/\pi_s\right) = \exp\left(-\frac{N-n}{n}\right)$$

and does not depend on x. This type of result is of no use to us since all sample uniques will be assigned equal risks.

Incorporating information from the population level here in terms of marginal counts was seen to sharpen the individual risk measures and significantly improve the stability. For an agency considering releasing a microdata set it seems prudent to consider previous releases from the data when assessing disclosure risks, e.g. in the form of marginal counts or simple two-way tables, as a cunning intruder very well might do.

As for the computational effort we found that the PiG based risk measure on average required only approximately 1/1900 of the time used for calculating the PLN measures. Although we do not claim that the programming is optimal, the difference still provides an indication of the time saved by using the PiG based measure or the simplified risk measure in (8).

7 Acknowledgments

The author wishes to thank Boris Lorenc, Dep. of Statistics, Stockholm University, for his valuable help with the programming, and Ingegerd Jansson, Riksförsäkringsverket (National Social Insurance Board), for supplying the data from the Store database used in the examples. The author is also grateful for the comments and suggestions of Professor Daniel Thorburn, Dep. of Statistics, Stockholm University, in the preparation of this paper.

A Figures and Tables

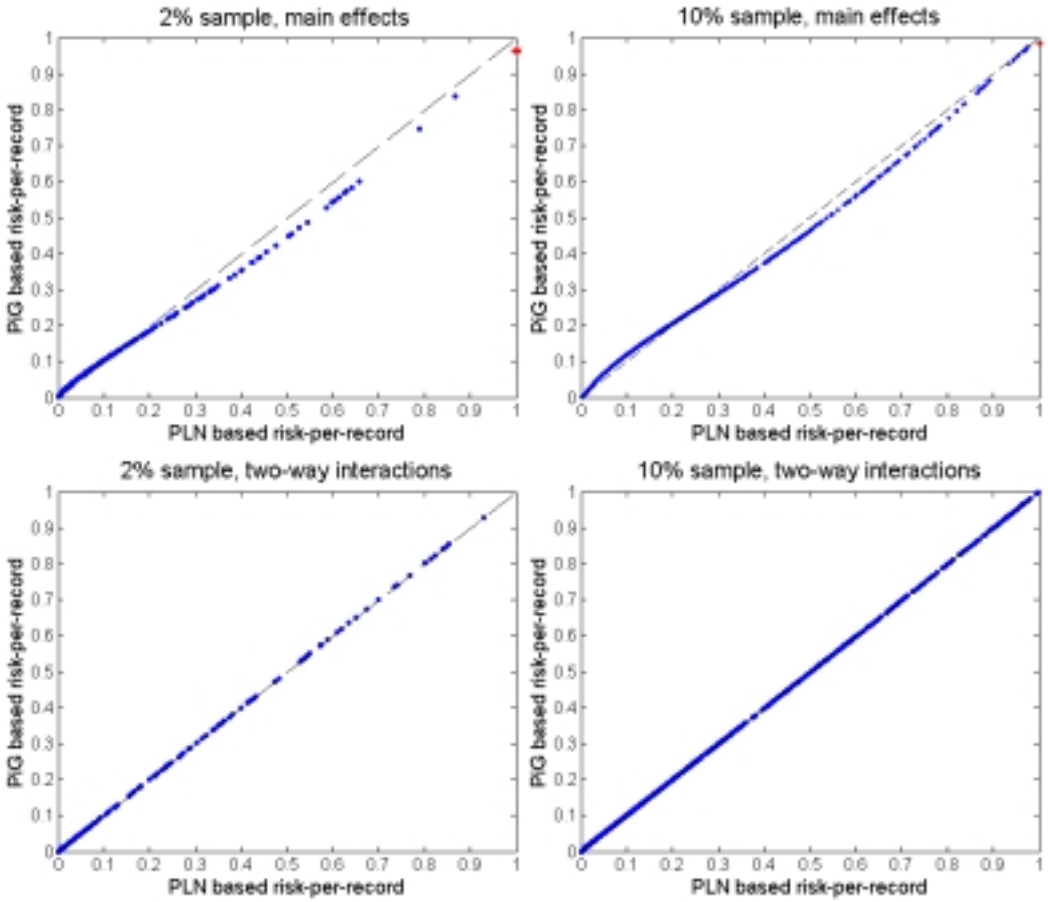

Figure 1: PiG based versus PLN based risk-per-record measures. The (*) in the upper left corners of the main effects only models indicate that the corresponding unit's PLN based risk measure is larger than one.

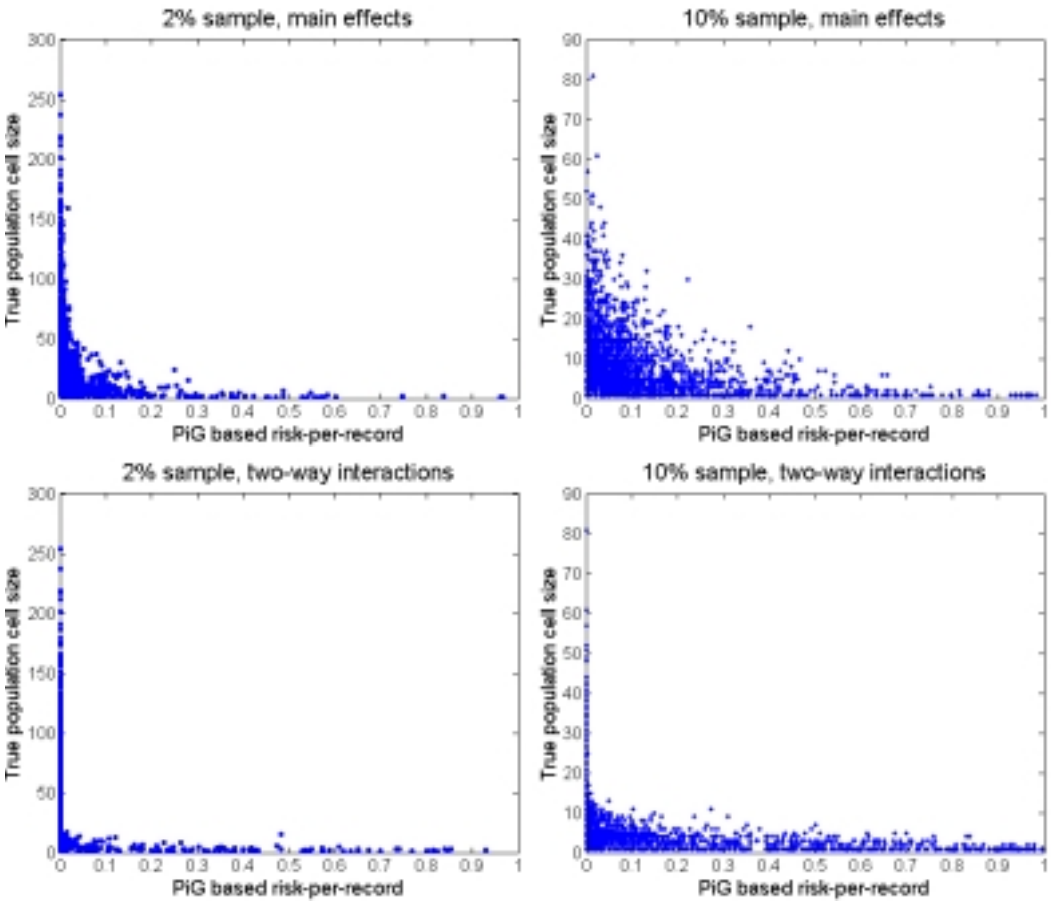

Figure 2: Population cell size versus PiG based risk measure.

Table 1: Distribution of population and sample cell sizes.

	Population		2% sample		10% sample
F_x	frequency	f_x	frequency	f_x	frequency
0	69,185	0	78,028	0	74,889
1	2,607	1	1,560	1	2,304
2	1,217	2	493	2	955
3	807	3	243	3	589
4	593	4	107	4	407
5	458	5	75	5	286
6	410	6	44	6	212
7	336	7	25	7	147
8	307	8	17	8	122
9	302	9	11	9	85
10	245	10	6	10	77
⋮	⋮	⋮	⋮	⋮	⋮
1,426	1	37	1	154	1

Table 2: Percentage of population uniques by range of risk measures, main effects model, 2% sample. The (*) indicates the number of units with risk measures larger than one.

Main effects model, 2% sample						
	PiG, $\hat{\tau} = 2.852$			PLN, $\hat{\sigma}^2 = 1.154$		
Range of risk measures	# sample uniques	% pop. uniques	% pop. pairs	# sample uniques	% pop. uniques	% pop. pairs
0 - 0.1	1,444	1.0	1.9	1,447	1.0	1.9
0.1 - 0.2	63	11.1	7.9	56	10.7	7.1
0.2 - 0.3	20	20.0	20.0	15	6.7	20.0
0.3 - 0.4	12	33.3	33.3	14	42.9	28.6
0.4 - 0.5	7	42.9	42.9	9	22.2	44.4
0.5 - 0.6	7	71.4	0	7	71.4	14.3
0.6 - 0.7	1	100.0	0	6	66.7	0
0.7 - 0.8	2	50.0	50.0	2	50.0	50.0
0.8 - 0.9	1	0	100.0	1	0	100.0
0.9 - 1.0	3	100.0	0	0 + 3*	100.0	0
Total	1,560	2.7%	2.9%	1,560	2.7%	2.9%

Table 3: Percentage of population uniques by range of risk measures, main effects model, 10% sample. The (*) indicates the number of units with risk measures larger than one.

Main effects model, 10% sample						
	PiG, $\hat{\tau} = 2.844$			PLN, $\hat{\sigma}^2 = 1.271$		
Range of risk measures	# sample uniques	% pop. uniques	% pop. pairs	# sample uniques	% pop. uniques	% pop. pairs
0 - 0.1	1,524	3.0	5.7	1,614	3.7	5.9
0.1 - 0.2	381	14.4	12.1	299	14.0	13.4
0.2 - 0.3	173	24.9	9.8	153	24.8	9.8
0.3 - 0.4	83	37.4	20.5	75	38.7	17.3
0.4 - 0.5	52	42.3	19.2	59	33.9	17.0
0.5 - 0.6	25	56.0	16.0	24	54.2	20.8
0.6 - 0.7	27	51.8	33.3	33	51.5	30.3
0.7 - 0.8	15	86.7	6.7	20	80.0	15.0
0.8 - 0.9	13	76.9	23.1	16	81.2	18.8
0.9 - 1.0	11	100.0	0	10 + 1*	100.0	0
Total	2,304	11.2%	8.4%	2,304	11.2%	8.4%

Table 4: Percentage of population uniques by range of risk measures, all two-way interactions model, 2% sample. Simplified indicates that the estimates of τ and σ^2 were negative and the simplified measure used.

All two-way interactions model, 2% sample						
	PiG, simplified			PLN, simplified		
Range of risk measures	# sample uniques	% pop. uniques	% pop. pairs	# sample uniques	% pop. uniques	% pop. pairs
0 - 0.1	1,460	0.9	1.5	1,460	0.9	1.5
0.1 - 0.2	25	12.0	12.0	25	12.0	12.0
0.2 - 0.3	20	35.0	20.0	20	35.0	20.0
0.3 - 0.4	15	26.7	26.7	15	26.7	26.7
0.4 - 0.5	10	20.0	30.0	10	20.0	30.0
0.5 - 0.6	12	41.7	41.7	12	41.7	41.7
0.6 - 0.7	6	50.0	16.7	6	50.0	16.7
0.7 - 0.8	4	50.0	0	4	50.0	0
0.8 - 0.9	7	42.9	28.6	7	42.9	28.6
0.9 - 1.0	1	0	100.0	1	0	100.0
Total	1,560	2.7%	2.9%	1,560	2.7%	2.9%

Table 5: Percentage of population uniques by range of risk measures, all two-way interactions model, 10% sample. Simplified indicates that the estimate of σ^2 was negative and the simplified measure used.

All two-way interactions model, 10% sample						
	PiG, $\hat{\tau} = 0.0118$			PLN, simplified		
Range of risk measures	# sample uniques	% pop. uniques	% pop. pairs	# sample uniques	% pop. uniques	% pop. pairs
0 - 0.1	1,737	2.3	4.5	1,740	2.3	4.5
0.1 - 0.2	153	15.0	17.0	150	15.3	17.3
0.2 - 0.3	90	31.1	20.0	87	29.9	20.7
0.3 - 0.4	60	30.0	23.3	61	32.8	19.7
0.4 - 0.5	66	42.4	22.7	66	40.9	25.8
0.5 - 0.6	52	46.2	30.8	53	45.3	30.2
0.6 - 0.7	39	43.6	25.6	40	45.0	25.0
0.7 - 0.8	29	55.2	20.7	28	53.6	21.4
0.8 - 0.9	35	80.0	14.3	35	80.0	14.3
0.9 - 1.0	43	86.1	13.9	44	86.4	13.6
Total	2,304	11.2%	8.4%	2,304	11.2%	8.4%

Table 6: Percentage of population uniques by range of PiG based risk measure using known population level main marginal counts and the ME model.

Main effects model using known population margins, PiG						
	2% sample, $\hat{\tau} = 2.842$			10% sample, $\hat{\tau} = 2.852$		
Range of risk measures	# sample uniques	% pop. uniques	% pop. pairs	# sample uniques	% pop. uniques	% pop. pairs
0 - 0.1	1,437	0.8	2.0	1,524	3.1	5.8
0.1 - 0.2	65	12.3	6.2	383	14.4	11.5
0.2 - 0.3	22	22.7	13.6	176	24.4	10.2
0.3 - 0.4	14	21.4	42.9	85	35.3	20.0
0.4 - 0.5	8	50.0	25.0	51	49.0	15.7
0.5 - 0.6	8	87.5	0	25	48.0	32.0
0.6 - 0.7	1	0	0	26	65.4	26.9
0.7 - 0.8	2	0	100.0	10	80.0	10.0
0.8 - 0.9	0	-	-	9	88.9	11.1
0.9 - 1.0	3	100.0	0	15	93.3	6.7
Total	1,560	2.7%	2.9%	2,304	11.2%	8.4%

Table 7: Percentage of population uniques by range of PiG based risk measure using known population level two-way marginal counts and the TW model. Simplified indicates that the estimate of τ was negative and the simplified measure used.

Known two-way margins and all two-way interactions model, PiG						
	2% sample, $\hat{\tau} = 0.0129$			10% sample, simplified		
Range of risk measures	# sample uniques	% pop. uniques	% pop. pairs	# sample uniques	% pop. uniques	% pop. pairs
0 - 0.1	1,439	0.3	0.8	1,709	1.4	3.5
0.1 - 0.2	36	8.3	16.7	146	13.7	19.9
0.2 - 0.3	18	16.7	33.3	84	20.2	25.0
0.3 - 0.4	15	20.0	33.3	66	30.3	25.8
0.4 - 0.5	10	40.0	40.0	59	28.8	35.6
0.5 - 0.6	8	37.5	50.0	56	46.4	28.6
0.6 - 0.7	11	45.4	27.3	43	60.5	18.6
0.7 - 0.8	8	62.5	25.0	43	62.8	23.3
0.8 - 0.9	7	71.4	28.6	42	76.2	16.7
0.9 - 1.0	8	87.5	12.5	56	91.1	8.9
Total	1,560	2.7%	2.9%	2,304	11.2%	8.4%

References

[1] Bethlehem, J.G., Keller, W.J. and Pannekoek, J. (1990) Disclosure Control of Microdata. *Journal of the American Statistical Association*, **85**, pp. 38-45.5-144.

[2] Benedetti, R., Franconi, L. and Piersimoni, F. (1999) Per-record Risk of Disclosure in Dependent Data. In *Statistical data protection - Proceedings of the Conference, Lisbon, 25 to 27 March 1998 - 1999 edition*, pp. 287-291.

[3] Carlson, M. (2002) Assessing Microdata Disclosure Risk Using the Poisson-Inverse Gaussian Distribution. *Statistics in Transition*, to appear.

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

[5] Christensen, R. (1997) *Log-Linear Models and Logistic Regression*, second edition. Monograph. New York: Springer-Verlag.

[6] Dalenius, T. (1977) Towards a Methodology For Statistical Disclosure Control. *Statistisk Tidskrift,* **5**, pp. 429-444.

[7] Dean, C., Lawless, J.F. and Willmot, G.E. (1989) A mixed Poisson-inverse-Gaussian regression Model. *The Canadian Journal of Statistics*, **17**, pp. 171-181.

[8] Domingo-Ferrer, J. (Ed) (2002) *Inference Control in Statistical Databases.* Monograph. Berlin: Springer.

[9] Doyle, P., Lane, J., Theeuwes, J.J.M., and Zayatz, L. (Eds.), (2001) *Confidentiality, Disclosure, and Data Access.* Monograph. Amsterdam: Elsevier.

[10] Duncan, G. and Lambert, D. (1989) The Risk of Disclosure for Microdata. *Journal of Business & Economic Statistics*, **7**, pp. 207-217.

[11] Duncan, G.T. and Pearson, R.W. (1991) Enhancing Access to Microdata while Protecting Confidentiality: Prospects for the Future. With discussion. *Statistical Science*, **6**, pp. 219-239.

[12] Elliot, M. (2002) Integrating File and Record Level Disclosure Risk Assessment. In J. Domingo-Ferrer (Ed.) *Inference Control in Statistical Databases*, pp.126-134. LNCS 2316, Heidelberg: Springer-Verlag.

[13] Elliot, M.J., Skinner, C.J. and Dale, A. (1998) Special Uniques, Random Uniques and Sticky populations: Some Counterintuitive Effects of Geographical Detail on Disclosure Risk. *Research in Official Statistics*, **1**, pp. 53-67.

[14] Fienberg, S.E. (1994) Conflicts Between the Needs for Access to Statistical Information and Demands for Confidentiality. *Journal of Official Statistics*, **10**, pp. 115-132.

[15] Fienberg, S.E. and Makov, U. (1998) Confidentiality, Uniqueness and Disclosure Limitation for Categorical Data. *Journal of Official Statistics*, **14**, pp. 385-397.

[16] Folks, J.L. and Chhikara, R.S. (1978) The Inverse Gaussian Distribution and Its Statistical Application - A Review. With discussion. *Journal of the Royal Statistical Society, Series B*, **40**, pp. 263-289.

[17] Frank, O. (1976) Individual Disclosures from Frequency Tables. In T. Dalenius and A. Klevmarken (eds.) *Personal Integrity and the Need for Data in the Social Sciences.* Swedish Council for Social Science Research, pp. 175-187.

[18] Frank, O. (1988) Designing Classifiers for Partial information Release, in H.H. Bock (editor) *Classification and related methods of data analysis : proceedings of the First Conference of the International Federation of Classification Societies,* pp*. 687-690.* New York: North-Holland.

[19] Grundy, P.M. (1951) The Expected Frequencies in a Sample of an Animal Population in which the Abundances of Species are Log-normally Distributed. Part 1. *Biometrika*, 38, pp. 427-434.

[20] Hoshino, N. (2001) Applying Pitman's Sampling Formula to Microdata Disclosure Risk Assessment. *Journal of Official Statistics*, **17**, pp. 499-520.

[21] Lambert, D. (1993) Measures of Disclosure Risk and Harm. *Journal of Official Statistics*, **9**, pp. 313-331.

[22] Marsh, C., Dale, A., and Skinner, C.J. (1994) Safe Data Versus Safe Settings: Access to Microdata from the British Census. *International Statistical Review*, 62 pp. 35-53.

[23] Matlab (2001) Matlab Version 6.1, Release 12.1, MathWorks Inc., http://www.mathworks.com.

[24] Ord, J.K. and Whitmore, G. (1986) The Poisson-Inverse Gaussian distribution as a model for species abundance, *Communications in Statistics - Theory and Methods*, **15**, pp. 853-871.

[25] Riksförsäkringsverket (2002) Store Database, Swedish National Social Insurance Board, http://www.rfv.se/english/index.htm.

[26] Samuels, S.M. (1998) A Bayesian, Species-Sampling-Inspired Approach to the Uniques Problem in Microdata Disclosure Risk Assessment. *Journal of Official Statistics*, **14**, pp. 373-383.

[27] Särndal, C.E., Swensson, B. and Wretman, J. (1992) *Model Assisted Survey Sampling*. Monograph. New York: Springer-Verlag.

[28] Sichel, H.S. (1974) On a Distribution Representing Sentence-length in Written Prose, *Journal of the Royal Statistical Society, Series A*, **137**, pp. 25-34.

[29] Sichel, H.S. (1975) On a Distribution Law for Word Frequencies, *Journal of the American Statistical Association*, **70**, pp. 542-547.

[30] Sichel, H.S. (1982) Repeat-buying and the Generalized Inverse Gaussian-Poisson Distribution, *Applied Statistics*, **31**, pp.193-204.

[31] Skinner, C.J. and Holmes, D.J. (1993) Modelling Population Uniqueness. In *Proceedings of the International Seminar on Statistical Confidentiality, Dublin, 8-10 September 1992*, pp. 175-199. Luxembourg: Office for Official Publications of the European Communities.

[32] Skinner, C.J. and Holmes, D.J. (1998) Estimating the Re-identification Risk Per Record. *Journal of Official Statistics*, **14**, pp. 361-372.

[33] Skinner, C., Marsh, C., Openshaw, S. and Wymer, C. (1994) Disclosure Control for Census Microdata, *Journal of Official Statistics*, **10**, pp. 31-51.

[34] Takemura, A. (1999) Some Superpopulation Models for Estimating the Number of Population Uniques. In *Statistical Data Protection - Proceedings of the Conference, Lisbon, 25 to 27 March 1998 - 1999 edition*, pp. 59-76. Luxembourg: Office for Official Publications of the European Communities.

[35] Takemura, A. (2001) Evaluation of per-record identification Risk by Additive Modeling of Interaction for Contingency Table Cell Probabilities. In *Proceedings, Invited Papers, The 53nd Session of The International Statistical Institute, August 22-29, 2001, Seoul, South Korea,* pp. 220-235. IASS.

[36] Willenborg, L. and de Waal, T. (1996) *Statistical Disclosure Control in Practice; Series: Lecture Notes in Statistics*, Vol. **111**. Monograph. Springer-Verlag, New York.

[37] Willenborg, L.C. and de Waal, T. (2000) *Elements of Statistical Disclosure Control; Series: Lecture Notes in Statistics*, Vol. **155**. Monograph. New York: Springer-Verlag.

[38] Willmot, G.E. (1987) The Poisson-inverse Gaussian Distribution as an Alternative to the Negative Binomial, *Scandinavian Actuarial Journal*, pp. 113-127.