Research Report

Department of Statistics

No. 2002:8

An Empirical Comparison of Some Methods for Disclosure Risk Assessment

Michael Carlson

Department of Statistics, Stockholm University, SE-106 91 Stockholm, Sweden

An Empirical Comparison of Some Methods for Disclosure Risk Assessment

Michael Carlson[*]

December 18, 2002

Abstract

With the release of public-use microdata files it is important to assess the risk of disclosing individual information. A measure of disclosure risk often considered in the literature is the proportion of unique records in the file that are also unique in the population. Various methods based on superpopulation models have been proposed for estimating this quantity using sample data. An empirical comparison of a selection of models applied to three real-life data sets is presented. The general conclusion is that no one model is uniformly best with respect to the risk measure used and that performance varies greatly between different types of data.

Keywords: Method evaluation; Statistical disclosure control; Superpopulation; Uniqueness.

1 Introduction

A provider of statistical microdata, typically a national statistical agency, must consider the rights of the respondents, both individuals and organizations, when releasing statistical data such as e.g. public-use microdata files or large complex tables. These rights, often prescribed by law, include the protection against unnecessary exposure, i.e. disclosure of confidential information about the respondents. It is therefore essential that the provider can assess the disclosure risk involved with the release of detailed data. Up to date a wide variety of methods of assessing statistical disclosure risk by means of the so-called uniqueness concept have been proposed, many based

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

on superpopulation assumptions although non-parametric approaches have also been considered e.g. in Greenberg and Zayatz (1993) and more recently in Elliot (2002) and Skinner and Elliot (2002).

This paper is a first small scale investigation of the performance of superpopulation models for disclosure risk assessment and aims at providing some further empirical results, which may generally be of guidance in comparing different methods. In evaluating the various methods it is important to apply them to data with varying properties and compare each method on the same appropriate criterion. The focus here is on disclosure risk on the file-level, i.e. measures of disclosure risk that pertain to the file or sample as a whole in contrast to per-record measures of risk. The risk measures are defined as the number of unique units in the population and as the proportion of unique records in a microdata file that are unique in the population. A selection of model-based methods are compared by means of simulation studies on three real-life data sets. The simulation study entails drawing samples of varying size from the data sets and for each sample computing an estimate of the risk measure defined. Similar comparative studies have been reported in among others Skinner and Holmes (1993), Chen and Keller-McNulty (1998), and Hoshino (2001).

This paper is organized as follows. The following section introduces some basic notation and describes the problem at hand. In section 3 we briefly discuss statistical disclosure risk as it is often defined in terms of uniqueness. The methods included in this study are described in section 4 and the data sets used are described in section 5. The results of the study are reported in section 6 and section 7 finalizes the paper with some concluding remarks.

2 Preliminaries

Consider a finite population U of size N from which a random sample without replacement $s \subseteq U$ of size $n \leqslant N$ is drawn. With each unit in U is associated the values of a number of discrete variables, $X_1, \ldots, X_q$ with $C_1, \ldots, C_q$ categories respectively. The cross-classification of these variables defines the discrete variable X with $\Pi C_i = C$ categories or *cells* and for simplicity we let the cells of X be labeled as $1, 2, \ldots, C$. The sample can be viewed as the public-use microdata file being considered for release containing records corresponding to individual respondents. It is well known that there remains the possibility that statistical disclosure could occur even when such a file has been anonymized by the removal of direct identifiers such as names, addresses and identity numbers, see e.g. Block and Olsson (1976) and Dalenius (1986). Respondents with a unique set of scores on the attributes are obviously at greater risk of being spotted in a re-identification process relative to other non-unique respondents.

Following e.g. Bethlehem et al. (1990) the X_i are termed *key variables*, X the *key* and the C different categories of X, the *key values*. Thus, the key divides the population into C subpopulations $U_i \subseteq U$ and by F_i we denote the number of units belonging to subpopulation U_i, i.e. the population frequency or size of cell i for $i = 1, \ldots, C$. The sample counterpart is denoted by f_i. Define T_j and its sample counterpart t_j as the number of cells of size j, i.e.

$$T_j = \#\left(i;\ F_i = j\right), \qquad j = 0, 1, \ldots, N$$

and

$$t_j = \#\left(i;\ f_i = j\right) \qquad j = 0, 1, \ldots, n$$

respectively. The T_j and t_j are usually termed cell size indices or frequencies of frequencies and correspond to the equivalence classes of Greenberg and Zayatz (1992). Of these quantities, C, N and n are usually fixed in the design, the f_i and t_j are observed from the sample and the F_i and T_j are assumed to be unknown.

Employing a superpopulation model implies that the individual population cell frequencies are realizations of a random variable from some suitable distribution, $\Pr\left(F_i = j\right) = P_j$, $j = 0, 1, 2 \ldots$. The goal is to model and estimate the population frequency structure, i.e. the T_j, based on sample information. If the distribution is well chosen and the parameters reliably estimated, it is then possible to reliably predict the number of cells of a certain size j in the population.

As the number of combinations of key variable values may be very large, it is inevitable that a significant number of cells will be empty simply by chance or because they are rare combinations. A further consideration deals with the fact that certain combinations may be logically impossible, such as married 4-year olds. Let S_0 denote the number of these so-called structural zeroes. If this number is known a priori it suffices to consider the non-structural zero cells in the modelling process. If S_0 is unknown or difficult to assess, a solution is to consider models which take on only positive valued integers, $j = 1, 2, 3 \ldots$ disregarding all empty cells. Alternatively the model can be extended to allow for a proportion of the cells to be zero with unit probability, as suggested by Skinner and Holmes (1993) and used in Carlson (2002a). The marginal distribution of the cell frequencies would then be given by

$$\Pr\left(F_i = j\right) = \theta I_{j=0} + \left(1 - \theta\right) P_j. \tag{1}$$

where $I_{(\cdot)}$ is the usual indicator function, θ is an additional parameter such that $0 \leqslant \theta < 1$ and P_j denotes a probability function.

An important issue is also the marginal distribution of the sample level cell frequencies, i.e. the distribution of the f_i under the sampling design and the assumed distribution of the population cell frequencies. Often the model and sampling designs considered result in sample distributions of the same form as the population level model, i.e. if $\Pr(F_i = j) = g(j; \theta)$ where g is a probability function and where θ is a set of parameters, then $\Pr(f_i = j) = g(j; \theta_s)$ where θ_s is some transformation of θ determined by the sampling design. Thus, the parameter θ_s is estimated from the sample data and then transformed to obtain an estimate of θ which in turn is used to predict the number of cells of size j in the population. See e.g. the discussion concerning sampling in Takemura (1999) and Hoshino (2001).

3 Risk assessment

The basic framework considered here is the same as presented by many authors, see e.g. Bethlehem et al. (1990), Paass (1988), Fienberg and Makov (1998), Skinner and Holmes (1998) and Skinner and Elliot (2002). Here we will only briefly review the basic ideas of assessing disclosure risk that build on the concept of uniqueness.

A *unique* is defined as an entity that has a unique set of values on the key variables. A unit that is unique in the population is referred to as a *population unique* whereas a unit that is unique in the sample is referred to as a *sample unique*. If a population unique is included in the sample it is necessarily also a sample unique but the converse does not hold; the occurrence of a sample unique does not imply that it is also a population unique. A population unique is obviously subjected to a greater risk of being exposed if it is included in the released data relative to other non-unique records.

A first measure of risk is the proportion or equivalently the number of population uniques, T_1. The idea is that if a perceived intruder could use the key to link an identifiable unit in the population to a record in the file and in addition knew that the unit was unique in the population then he could deduce that the link is correct. The proportion T_1/N is interpreted as the probability that a unit drawn at random from the population is population unique, assuming equal sampling probabilities across units.

However, it could be argued that an intruder will be more inclined to focus upon those records that are sample unique since it is only these that can by definition be population uniques. An alternative measure is thus the proportion of sample uniques that are also population uniques, i.e.

$$R = \frac{\#\,(\text{records that are Pop.Uniques } \textit{and} \text{ SampleUniques})}{\#\,(\text{records that are SampleUniques})}. \tag{2}$$

This proportion is interpreted as the probability that a sample unique is population unique. The intruder is thought of as drawing one sample unique at random from the set of sample uniques and then search through the population units at random until a matching population unit is found, see e.g. Skinner and Elliot (2002).

Let $t_{1,1}$ denote the nominator of (2). Under simple random sample without replacement or Bernoulli sampling with inclusion probability denoted by $\pi_s = n/N$, a first result is given by the expectation of $t_{1,1}$ given the value of T_1 which is

$$E\left(t_{1,1}\right) = \pi_s T_1 \tag{3}$$

since each population unique is equally likely to be included in the sample. Secondly, as shown by Greenberg and Zayatz (1992), the expected number of sample uniques t_1 given the sample size n and the T_j, is derived from

$$E\left(t_1 \mid n, T_1, T_2, \dots\right) = \sum_{j=1}^{N-n+1} \frac{\binom{N-j}{n-1}}{\binom{N}{n}} j T_j. \tag{4}$$

So, given that a unit is unique in the sample, the expected conditional probability that it is also unique in the population is approximately given by the ratio

$$\text{Risk} = E\left(R \mid n, T_1, T_2, \dots\right) \approx \frac{T_1/N}{E\left(t_1 \mid T_1, T_2, \dots\right)/n} \tag{5}$$

which can be used as a risk measure of any given data set. Usually it is assumed that the intruder will not know the true value of the F_i, and hence the T_j, since the microdata contains only a sample. By introducing a superpopulation model and viewing the population level cell frequencies as a realization of this generating model he may attach a probability distribution $g(j) = Pr(F_i = j)$ to the F_i, preferably defined through some parametric family and the conditional probability under the model would be given by

$$\text{Risk} = \Pr\left(F_i = 1 \mid f_i = 1\right). \tag{6}$$

By replacing the parameters for their estimates in (6) the matter of risk assessment is reduced to a matter of parameter estimation and prediction.

As a method of comparing different keys on the population level we consider also the so-called identifying force of the key. The *resolution* of a key, proposed by Block and Olsson (1976) and Bethlehem et al. (1990), is defined by

$$R(X) = \left[\sum_{i=1}^{C} \left(\frac{F_i}{N} \right)^2 \right]^{-1}$$

and equals the reciprocal of the probabilities that two random units selected with replacement from the population have the same key value. The resolution is compared to the size of the population; if all units are unique the resolution equals N. The identifying force is also expressible as the *entropy* of the key. The entropy of a random variable is defined as $-E\,(p_X \log p_X)$, where p_x is the probability function; for an introduction to information theory see e.g. Cover and Thomas (1991). Information theory and entropy measures have been adapted as a means of assessing disclosure risk in e.g. Frank (1978, 1983, 1988) Greenberg and Zayatz (1992) and Carlson (2002b). Given no knowledge of the key for a randomly selected unit from the population, the uncertainty of a correct link measured in terms of entropy is given by $\log(N)$. The reduction of the initial uncertainty after gaining knowledge of the key, is the entropy of the key, defined by

$$H(X) = -\sum_{i=1}^{C} \frac{F_i}{N} \log \frac{F_i}{N}.$$

Both the entropy and the resolution of the key provide measures of the identifying force of the key. Large values of entropy and resolution indicate that the key should be considered as dangerous. Note also that the identifying force indicates the average risk over all possible key values.

4 Superpopulation Models

In this section we give brief descriptions of the various superpopulation models that are included in the study. Three of the models are compound Poisson models, the negative binomial (NB), the Poisson-lognormal (PLN) and the Poisson-inverse Gaussian (PiG), where the cell frequencies are assumed to be generated independently from Poisson distributions with individual rates λ_i, $i = 1, \ldots, C$. 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_i. Given the N, C and the π_i the frequencies will follow a multinomial distribution and if the number of cells is large enough each cell frequency is approximately independently binomial with parameters N and π_i. Since the population size is usually quite large and the π_i small due to large C the Poisson distribution is used to approximate the binomial with $\lambda_i = N\pi_i$. To simplify the model further the λ_i are viewed as independent realizations of a continuous random variable Λ with a common probability density function (pdf) $g(\lambda)$. The number of cells C is usually quite large and this assumption will significantly reduce the number of parameters that need to be estimated. The specification of the mixing distribution $g(\lambda)$ is the crucial step and several different suggestions have been studied. The other two models are the logarithmic series distribution (LSD) and the Pitman's sampling formula. The LSD can be derived as a limiting case of the NB. The Pitman's model is a generalization of Ewen's sampling formula which is a conditional model of the LSD where the population size is fixed, see e.g. Hoshino and Takemura (1998). For more detailed descriptions of the respective models we refer to the references cited.

4.1 Negative Binomial (NB)

This negative binomial, or Poisson-gamma, was proposed in Bethlehem et al. (1990) as a possible model for disclosure risk assessment and was perhaps the first approach towards a superpopulation model. The probability mass function (pmf) of the cell frequencies is defined by

$$P_j = \frac{\Gamma(\alpha + j)}{\Gamma(\alpha)\Gamma(j+1)} \left(\frac{1}{\beta+1}\right)^{\alpha} \left(\frac{\beta}{\beta+1}\right)^{j}$$

for $j = 0, 1, 2, \ldots$ and where $\alpha, \beta > 0$ under the constraint that $\alpha = N/\beta(C - S_0)$. Assuming Bernoulli sampling, the cell sizes are also distributed as NB with parameters α and $\beta_s = \pi_s\beta$. This model has been noted to provide a poor fit to real-life data sets, see e.g. Skinner et al. (1994), and Chen and Keller-McNulty (1998). We included it in our investigation for purposes of comparison. Furthermore, we used an alternative moment based estimator that utilizes the sample mean and the proportion of uniques amongst the non-empty cells, described in e.g. Johnson et al. (1992, p. 226) and added the extra parameter θ as in (1). The expected number of population uniques under this model is given by

$$E(T_1) = C(1-\theta)\alpha\left(\frac{1}{\beta+1}\right)^{\alpha}\left(\frac{\beta}{\beta+1}\right)$$

and the risk measure in (6) simplifies to

$$Risk = \left(\frac{\pi_s\beta + 1}{\beta + 1}\right)^{\alpha+1}.$$

A variant of the NB model was proposed by Chen and Keller-McNulty (1998) who proposed a shifted negative binomial distribution based on their findings of typical cell size distributions in real-life data.

4.2 Logarithmic series distribution (LSD)

In many cases it has been noted that the α parameter of the NB-model tends to be very small in disclosure applications. In such cases it may be appropriate to consider instead the limiting distribution of a (zero-truncated) NB as $\alpha \to 0$ which results to Fisher's logarithmic series distribution, see e.g. the presentation given by Skinner and Holmes (1993). Writing $\beta = \phi/(1-\phi)$ the pmf is defined as

$$\Pr(F_i = j) = -\frac{\phi^j}{j \log(1-\phi)}$$

for $j = 1, 2, \ldots$ and where $0 < \phi < 1$. Assuming Bernoulli sampling results in the sample distribution also being LSD with parameter

$$\phi_s = \frac{\pi_s\phi}{1 - \phi(1-\pi_s)}.$$

The risk measure (5) is reduced to

$$Risk = 1 - \phi + \pi_s\phi$$

and the expected number of population uniques is given by

$$E(T_1) = N(1-\phi)$$

The parameter ϕ_s was estimated using ordinary maximum likelihood (ML) methods, see Skinner and Holmes (1993) for details.

4.3 Pitman's sampling formula

Hoshino (2001) proposed the Pitman sampling formula which is defined in terms of the cell size indices, i.e. the T_j. For each pair of real parameters α

and θ such that either $0 \leqslant \alpha < 1$ and $\theta > -\alpha$, or $\alpha < 0$ and $\theta = -m\alpha$ for some natural number m the Pitman model is defined by

$$P\left(T_1, \ldots, T_N\right) = N! \frac{\theta^{[U:\alpha]}}{\theta^{[N]}} \prod_{j=0}^{N} \left[\frac{(1-\alpha)^{[j-1]}}{j!} \right]^{T_j} \frac{1}{T_j!} \tag{7}$$

where $U = C - T_0$, the number of non-empty cells in the population,

$$\theta^{[N]} = \theta\left(\theta+1\right)\left(\theta+2\right)\cdots\left(\theta+\left(N-1\right)\right)$$

and

$$\theta^{[U:\alpha]} = \theta\left(\theta+\alpha\right)\left(\theta+2\alpha\right)\cdots\left(\theta+\left(U-1\right)\alpha\right).$$

If α equals zero, (7) amounts to the Ewen's model, studied by Samuels (1998). Assuming that $\alpha < 0$, and letting $\theta = -C\alpha$ it can be shown that (7) amounts to the multinomial-Dirichlet model investigated by e.g. Takemura (1999) and Hoshino and Takemura (1998). Note that the number of empty cells is not defined under this model. The sampling distribution under simple random sampling is also defined by (7) with N and U replaced by n and $u = C - t_0$, respectively. Using equation 26 in Hoshino (2001), we find that under the model, the expected number of population uniques is given by

$$E\left(T_1\right) = \frac{N\Gamma\left(\theta+\alpha+N-1\right)\Gamma\left(\theta+1\right)}{\Gamma\left(\theta+N\right)\Gamma\left(\theta+\alpha\right)}$$

The sample counterpart t_1 is estimated using n instead of N. The risk measure (5) is easily seen to simplify to

$$Risk = \frac{\Gamma\left(\theta+\alpha+N-1\right)\Gamma\left(\theta+n\right)}{\Gamma\left(\theta+\alpha+n-1\right)\Gamma\left(\theta+N\right)}$$

although Hoshino proposes using the observed number of sample uniques rather then the predicted value. Using the predicted ensures that it is consistent insofar that if the entire population is sampled, the measure equals 1. In this study we consider ML estimation, see Hoshino (2001) for details.

4.4 Poisson-lognormal (PLN)

Skinner and Holmes (1993) proposed the Poisson-lognormal distribution and also provide some heuristic justification for this model. The pmf of the population cell frequencies is defined by

$$P_j = \int_0^\infty \frac{\lambda^j e^{-\lambda}}{j!} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(\log\lambda-\mu\right)^2}{2\sigma^2}\right) d\lambda,$$

for $j = 1, 2, \ldots$ and where $\sigma^2 > 0$. The distribution of the sample cell frequencies under Bernoulli sampling is also PLN with parameters $\mu_s = \mu + \log \pi_s$ and σ^2. In our study we used the zero-truncated approach as described by Skinner and Holmes, using the extra parameter θ as in (1). As the lognormal distribution is not expressible in closed form, the calculation of the probabilities, and hence the risk measure (6), requires numerical integration methods. Estimation may result in a heavy computational burden but some effort is spared by censoring or truncating the likelihood above some suitable threshold value as discussed by Skinner and Holmes and by Bulmer (1974). In our study we tested different values for truncation as reported in the following sections. The numeric integration procedure used is described in Carlson (2002) and checked against the values tabulated in Grundy (1951).

4.5 Poisson-inverse Gaussian (PiG)

This model was suggested as a possible model for assessing disclosure risk in Carlson (2002a). The pmf of the PiG takes the form

$$P_j = \frac{\sqrt{\eta}}{j!} \left(\frac{\mu}{\eta}\right)^j \frac{K_{j-1/2}\left(\mu\eta/\tau\right)}{K_{-1/2}\left(\mu/\tau\right)}$$

for $j = 0, 1, \ldots$, $\mu, \tau > 0$ and where $\eta = \sqrt{1 + 2\tau}$. $K_\gamma\left(z\right)$ denotes a modified Bessel function of the third kind of order γ and argument z, see e.g. Carlson (2002) for details and references. In our study we used the zero-truncated approach as described previously, using the extra parameter θ as in (1). It is easily seen that under Bernoulli sampling the f_i are distributed as PiG with parameters $\mu_s = \pi_s\mu$ and $\tau_s = \pi_s\tau$, observing that $\eta_s = \sqrt{1 + 2\tau_s}$. The risk measure (6) simplifies to

$$Risk = \frac{\eta_s}{\eta} \exp\left[\frac{\mu}{\tau}\left(\eta_s - \eta\right)\right]$$

and the expected number of population uniques is given by

$$E\left(T_1\right) = C\left(1 - \theta\right) \frac{\mu}{\eta} \exp\left[\frac{\mu}{\tau}\left(1 - \eta\right)\right]$$

Various estimation methods are described in Carlson (2002a) of which two are considered in this study: ML and fitting the conditional probabilities $\Pr\left(f_i = 1 \mid f_i > 0\right)$ and $\Pr\left(f_i = 2 \mid f_i > 0\right)$ to the observed values of t_1 and t_2 relative the number of non-empty cells in the sample. I.e. one finds the solution to the pair of equations

$$\begin{cases} \dfrac{t_1}{C - t_0} = \Pr\left(f_i = 1 \mid f_i > 0\right) \\ \dfrac{t_2}{C - t_0} = \Pr\left(f_i = 2 \mid f_i > 0\right) \end{cases}$$

This is the method of estimation proposed by Chen and Keller-McNulty (1998) for their model and we will denote it in the following by PF12. Some further notes on this method of estimation are given in the appendix.

5 The Data Sets

A population consisting of individuals aged 18-65 residing in three counties in the southern part of Sweden was compiled from the Store database (Riksförsäkringsverket, 2002), managed by the Swedish National Social Insurance Board. After removing individuals for which the marital status was unknown (code = 8), the total population size was $N = 268,607$. Two sets of key variables were used as listed in Table 1 where the number of levels of each variable are given. The data sets corresponding to the two sets of key variables are denoted RFV5 and RFV7, respectively[1]. The smaller set RFV5, with five key variables, is a subset of those in RFV7.

A second population was compiled by Statistics Sweden, consisting of $N = 160,536$ individuals aged 20-65 living in one county in the central part of Sweden. The data for this set, which we denote by SCB7 originates from the 1990 Swedish census. The variable *marital status* has since been redefined which explains the change in number of levels as compared to the RFV sets. Furthermore, the variable *income* is in 50,000 SEK bands for the RFV data, the highest level also being the topcode, whereas for the SCB7 set, *income* is given in 10,000 SEK bands. For the RFV sets, the variable *citizenship* has three levels (Swedish, EU or Other) whereas for the SCB7 set, it takes on only two (Swedish or Foreign).

Some basic characteristics of the three data sets are given in table 1 and the cell size distributions of each set are plotted in Figure 1. We note the following differences between the sets: (a) the number combinations on the key variables ranges from small to large compared to the respective population size (b) the number of non-empty cells ranges from approximately 2 to 18 percent of the total number of cells, (c) the number of population uniques

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

Table 1: Description of the data sets and the key variables used in each.

	RFV5	RFV7	SCB7
Age	48	48	46
Sex	2	2	2
Marital status	7	7	10
Children	2	2	2
County	-	3	(1)
Municipality	-	-	6
Income	20	20	176
Citizenship	-	3	2
Total no. of cells, C	26,880	241,920	1,943,040
Non-empty cells, C-T_0	4,921	15,290	39,822
Population uniques, T_1	934	5,055	19,273
Largest cell size	2,741	1,389	140
Population size, N	268,607	268,607	160,536
Key resolution, R(X)	549.0	1410	8506
Key entropy, H(X)	7.150	12.50	9.802

ranges from approximately 0.3 to 12 percent of the population size, and from 18 to 48 percent of the number of non-empty cells, and (d) the right-hand tails of the RFV sets appear longer and heavier than the SCB7 set. This is further indicated by the largest observed cell sizes of each set.

The identifying forces of the respective keys, as measured by resolution and entropy, were calculated and we note that, not surprisingly, the SCB7 set possesses the most dangerous key. The number of occupied key values for this set is more than 2.5 times that of the RFV7 set and more than 8 times that of RFV5. Although the number of possible key values differs even more this will have no effect per se on the respective measures, as the contribution from empty cells is nil. The measures should be compared to N and $\log(N)$ respectively.

Given the population frequency structure it is possible to calculate the disclosure risk as measured by (5). This was done for the three sets and the results are shown in Figure 2 which shows the relationship between (5) and the sampling fraction π_s. When $\pi_s = 1$, then the risk is obviously 1. As previously discussed by St-Cyr (1998) we note that the relationship is concave and that the concavity increases with the number of key variables as seen when we compare RFV5 and RFV7. The set SCB7 is the most concave of the three which agrees with the measurements of identifying force above. The expectations of the number of population uniques in the sample, the number of sample uniques and the risk ratio as defined in (3), (4) and (5) are given in Table 2 for the sample fractions used in the simulation study. The goal is to obtain a risk measure that mimics this behavior.

6 Results of the Simulation Study

To evaluate the models a simulation study was conducted. Samples were drawn from each data set by simple random sampling without replacement, using varying sampling fractions, π_s. For the RFV sets $\pi_s = 1, 2, 5$ and 10% samples were drawn, and for the SCB7 set $\pi_s = 2.5, 5$, 10 and 25%. The reason for using a different set of sampling fractions for the SCB7 set was governed by the differences in the frequency structure compared to the RFV sets. E.g. the largest cell size in the SCB7 set is 140, compared to the 2,741 and 1,389 respectively for the RFV sets, and it is easily seen that the expected frequency structure of a 1% sample from the SCB7 set would hold only very small cell sizes leading to considerable difficulties when estimating model parameters.

For each sample drawn the parameters of each model were estimated as described in the preceding section. From the estimates the number of population uniques, T_1, and the risk measure as defined in (6), were predicted as well. Furthermore, for each sample the number of sample uniques, t_1, and the number of population uniques falling in the sample, $t_{1,1}$, were recorded for reference. From these we calculated for each sample the true ratio defined in (2). The means and standard deviations of these observed ratios are given in Table 2 together with the theoretically derived expected ratios, as described in the preceding section. It is clearly seen that the observed and the theoretically derived ratios agree. This is the risk measure that the model based approaches intend to predict. The means and standard deviations of the predicted risk measures and number of population uniques from estimating the models are given in Tables 3 and 4 for each simulation setting (data set and sampling fraction). The means and standard deviations of the model parameters are given Tables 5-10.

The NB model consistently underestimates the risk indicators in all simulation settings. A first indicator of the performance is that all the predicted values of the risk ratio lie below the corresponding sampling fraction and do not exhibit the concave which the concavity behavior seen in Figure 2. The extension of the model by including the mixture parameter θ as defined in (1) and the alternative zero-truncated moment estimators seem not to improve the results of this model, comparing the results with previous studies.

The LSD model yields a better performance compared to the NB; the

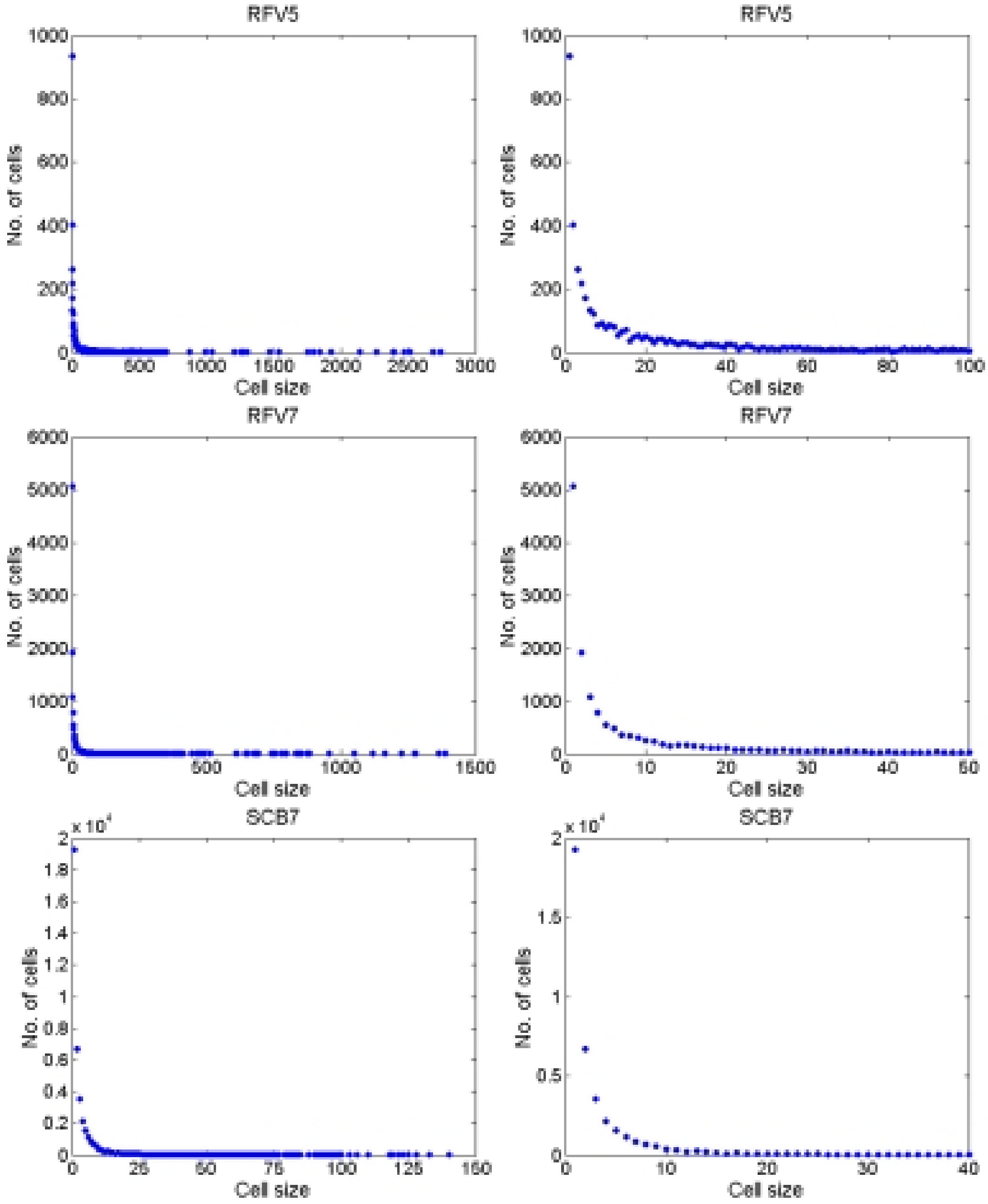

Figure 1: Population cell size distributions of the example data sets. The displays to the left show the entire distributions and the displays to the right show the details of the left-hand-side of the respective distributions. The frequencies of empty cells, i.e. T_0, are omitted in all cases.

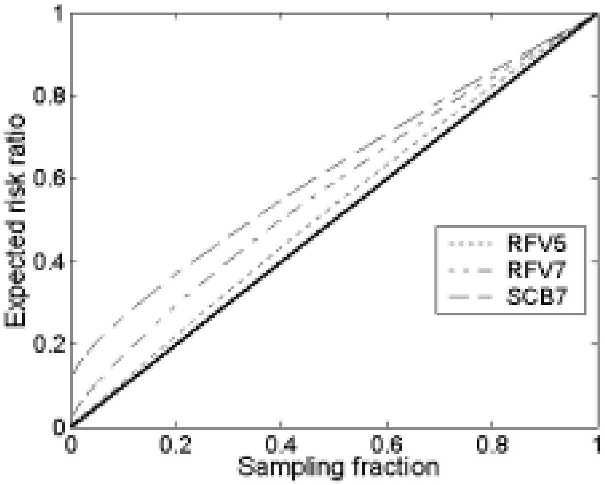

Figure 2: Theoretically derived expected risk measures for the example data. The solid line is for reference.

risk ratio and number of population for this model exhibits the concave behavior although they are slightly underestimating the true values as are the estimates of T_1. A possible reason is that the LSD is a single parameter distribution and may be less sensitive to model misspecification. We note also that both the NB and LSD models are quite stable yielding the lowest levels of variation.

Pitman's sampling formula is the only model that consistently overestimates the risk measure and the number of population uniques for these data sets, although the results improve with increasing sampling fraction. Also, the measures are not as stable as for the NB or LSD, showing a larger degree of variation. Overestimation is not as severe a problem as underestimation with respect to disclosure risk assessment, as it provides conservative indications of the risks. On the other hand if the indicators lead us to believe that a data material is too unsafe, we may apply disclosure limiting techniques to the data, and thereby unnecessarily reduce its analytical value.

As mentioned above we used the right-truncated approach for estimating the PLN model parameters, i.e. the likelihood was right-truncated and all values above a threshold value m were disregarded in the estimation process. As it is difficult to assess suitable threshold values a priori, we simply estimated the models for a small selection of values in each configuration and oppoturnisticly chose the value exhibiting the best performance with respect to bias in the risk indicators. Unfortunately the thresholds selected varied between the configurations, (5, 10, 20) or (10, 20, 30), and the results are therefore perhaps not always comparable. However, it was found that the best results were obtained with the lowest threshold values, save the SCB7 25 percent setting were $m = 10$ was superior to $m = 5$. With PLN model the risk indicators are underestimated for the RFV5 set, improved with the RFV7 set and with the SCB7 set the PLN performs among the better. We note also that the PLN failed to converge in 42 cases in the SCB7 2.5% setting. The results show that the PLN does not perform well with the RFV5 sets with estimated means on average lying between the NB and LSD. The results are improved for the RFV7 set although still underestimating the true values. For the SCB set on the other hand the performance is among the better, only slightly underestimating the true values, and with standard errors comparable to those of the Pitman and PiG.

There appears to be some problems with the PiG model and the methods of estimation used. For the RFV5 set the PiG model seriously underestimates both the risk measure and the number of populations uniques. For the RFV7 set the PiG model still underestimates these indicators although the results are improved but with the SCB7 set the PiG model is among the better, especially for the PF12 estimator with respect to bias, and standard errors comparable to the PLN and Pitman models. However, the PiG model did not converge in any of the cases when 2.5 and 5 percent samples were drawn from this set. Also in the RFV5 10% setting did the PF12 procedure fail to produce estimates. If the problem depends on a poor model or simply on the numerical procedures used in the study remains to be investigated however. We note also that the PiG failed to converge in 8, 8 and 36 cases for the RFV5 5%, RFV7 1% and SCB7 10% settings respectively.

7 Remarks

The scale of the present study comprised only three data sets of which two originate from the same source, and it is difficult to draw any general conclusions as to which models yields the best result. However, for these particular data sets we may conclude from the results that none of the models perform uniformly best and that the performance varies greatly between the different sets of data. The first observation is that apparently none of the models, except the LSD model, adapt well to the RFV5. For the RFV7 set the overall performance of all models is improved although none are quite satisfactory. With the SCB7 set the performance of the Pitman, PLN and PiG appear satisfactory at least for sample fractions at 10 and 25 percent.

A future prospect is to extend the present study with above all the inclusion of more diverse data sets. The three data sets used in the present study do not cover the entire range of diversity that might be encountered in practice. The studies should include sets from larger populations, a wider range of sampling fractions and a larger collection of keys. Other models and alternative estimation procedures may also be included in future studies. This would hopefully enable us to draw more extensive conclusions as to what type of models and estimation procedure is best suited for a particular data set and how to determine from the characteristics of the data

It is interesting to note that for this set the two procedures using the least amount of information from the sample, i.e. the right-truncated PLN and the PiG PF12, show the best performance with respect to bias. This seems to agree with the argument that in applications to disclosure control, a lack of fit in the right hand tail is not likely to be as critical as the left hand tail which may be considered more crucial since only cells belonging to t_0 and t_1 can by definition contain population uniques, see e.g. Skinner and Holmes (1993) and Chen and Keller-McNulty (1998). On the other hand these two models did not work at all for the RFV5 set which appears to agree with the argument presented in Hoshino (2001, p. 509) that it is better to utilize the whole information of the sample.

8 Acknowledgments

The author would like to thank Professor Daniel Thorburn, Dep. of Statistics, Stockholm University, for his valuable suggestions in the preparation of this paper. The author is also greatly thankful to Ingegerd Jansson, Riksförsäkringsverket (National Social Insurance Board), and Hans Block, Statistics Sweden, for supplying the data sets used.

A Tables

Table 2: Expected values (exp) of $t_{1,1}$, t_1 and $R \times 100$, simulated means and their standard errors, based on 1,000 simulations.

		$t_{1,1}$			t_1			$t_{1,1}/t_1 \times 100$		
Set	π_s	exp	mean	sd	exp	mean	sd	exp	mean	sd
RFV5	1%	9.34	9.16	(2.93)	632	633	(21.7)	1.48	1.45	(0.460)
	2%	18.7	18.6	(4.14)	739	738	(22.7)	2.53	2.52	(0.554)
	5%	46.7	46.4	(6.68)	823	823	(24.4)	5.68	5.63	(0.781)
	10%	93.4	92.8	(8.93)	842	841	(23.2)	11.1	11.0	(0.992)
RFV7	1%	50.6	50.5	(7.11)	1205	1206	(29.3)	4.20	4.19	(0.571)
	2%	101	101	(9.63)	1685	1685	(33.9)	6.00	6.01	(0.555)
	5%	253	253	(15.1)	2374	2375	(42.2)	10.6	10.6	(0.607)
	10%	506	506	(20.8)	2897	2893	(46.0)	17.4	17.5	(0.637)
SCB7	2.5%	482	481	(20.0)	2879	2877	(36.2)	16.7	16.7	(0.662)
	5%	964	965	(28.5)	4693	4698	(52.2)	20.5	20.5	(0.555)
	10%	1927	1926	(39.5)	7213	7214	(67.9)	26.7	26.7	(0.478)
	25%	4818	4816	(57.2)	11563	11563	(84.1)	41.7	41.6	(0.381)

Table 3: Predicted risk measures under the different models, simulated means and their standard errors, based on 1,000 simulations, save PLN SCB 2.5% where 42 cases failed to converge, and PiG pf12, RFV5 5%, RFV7 1% and SCB7 10% where 8, 8 and 36 cases failed to converge. The right-truncation threshold for the PLN models were chosen as indicated..

		$\hat{R} \times 100$						
Set	π_s	NegBin			LSD		Pitman	
RFV5	1%	0.393	(0.0320)		1.28	(0.00943)	3.37	(0.559)
	2%	0.953	(0.0416)		2.29	(0.00618)	4.19	(0.399)
	5%	3.05	(0.0546)		5.29	(0.00402)	7.22	(0.327)
	10%	7.11	(0.0590)		10.3	(0.00299)	12.12	(0.286)
RFV7	1%	0.459	(0.0361)		1.72	(0.0286)	8.94	(1.32)
	2%	1.08	(0.0428)		2.77	(0.0188)	9.78	(0.804)
	5%	3.35	(0.0517)		5.85	(0.0121)	13.8	(0.525)
	10%	7.65	(0.0536)		10.9	(0.00898)	19.8	(0.392)
SCB7	2.5%	1.16	(0.0844)		8.07	(0.233)	18.9	(3.04)
	5%	3.02	(0.0946)		10.8	(0.139)	24.6	(1.60)
	10%	7.26	(0.0876)		16.0	(0.0860)	32.6	(0.854)
	25%	21.3	(0.0668)		30.8	(0.0425)	47.7	(0.385)
Set	π_s	PLN			PiG, ML		PiG, pf12	
RFV5	1%	0.594	(0.759)	5	0.0727	(0.0530)	0.107	(0.169)
	2%	1.33	(1.07)	5	0.108	(0.0401)	0.242	(0.212)
	5%	3.40	(1.01)	10	0.414	(0.0684)	1.12	(0.442)
	10%	7.38	(1.38)	10	1.39	(0.117)	-	-
RFV7	1	0.998	(0.737)	5	3.95	(1.99)	1.97	(1.68)
	2	1.66	(0.519)	10	3.15	(0.862)	2.58	(1.32)
	5	6.58	(1.04)	10	4.88	(0.568)	5.32	(1.21)
	10	12.89	(1.18)	10	8.61	(0.493)	10.7	(1.28)
SCB7	2.5	8.36	(2.30)	5	-	-	-	-
	5	17.0	(2.30)	5	-	-	-	-
	10	26.1	(2.48)	5	30.7	(1.53)	30.2	(1.94)
	25	40.8	(0.737)	10	44.3	(0.630)	41.4	(1.05)

Table 4: Predicted no. of population uniques, T_1, under the different models, simulated means and their standard errors, based on 1,000 simulations, save PLN SCB 2.5% where 42 cases failed to converge, and PiG pf12, RFV5 5%, RFV7 1% and SCB7 10% where 8, 8 and 36 cases failed to converge. The right-truncation threshold for the PLN models were chosen as indicated.

		$\hat{T}_1$						
Set	π_s	NegBin			LSD		Pitman	
RFV5	1%	141	(11.5)		765	(25.6)	2181	(416)
	2%	201	(9.79)		794	(16.9)	1594	(188)
	5%	292	(7.58)		823	(11.4)	1237	(83.1)
	10%	361	(6.13)		834	(8.93)	1074	(46.4)
RFV7	1%	305	(21.8)		1940	(77.5)	10871	(1777)
	2%	486	(18.4)		2124	(51.6)	8331	(803)
	5%	808	(15.2)		2396	(34.2)	6666	(347)
	10%	1102	(13.2)		2613	(26.8)	5877	(191)
SCB7	2.5%	809	(53.6)		9164	(383)	21806	(3610)
	5%	1662	(47.7)		9775	(235)	23077	(1648)
	10%	2897	(38.8)		10677	(153)	23486	(778)
	25%	5026	(32.0)		12368	(90.9)	22241	(313)
Set	π_s	PLN			PiG, ML		PiG, pf12	
RFV5	1%	323	(357)	5	46.9	(35.8)	68.9	(112)
	2%	396	(264)	5	40.1	(16.0)	90.6	(82.1)
	5v	492	(125)	10	66.6	(12.8)	186	(77.3)
	10%	531	(67.5)	10	110	(12.0)	-	-
RFV7	1%	1199	(887)	5	4807	(2505)	2396	(2074)
	2%	1391	(444)	10	2668	(769)	2182	(1147)
	5%	3024	(487)	10	2322	(306)	2533	(607)
	10%	3539	(313)	10	2479	(177)	3102	(405)
SCB7	2.5%	9600	(2676)	5	-	-	-	-
	5%	15761	(2170)	5	-	-	-	-
	10%	18515	(1729)	5	22161	(1247)	21793	(1501)
	25%	18722	(411)	10	20530	(417)	19140	(576)

Table 5: Parameter estimates for zero-truncated negative binomial model, moment estimator, simulated means and standard errors, based on 1,000 simulations.

Set	π_s	$\hat{\alpha}$		$\hat{\beta}_s$		$\hat{\theta}$	
RFV5	1%	0.236	(0.0208)	7.88	(0.637)	0.893	(0.00547)
	2%	0.211	(0.0125)	13.66	(0.744)	0.860	(0.00525)
	5%	0.177	(0.00643)	30.32	(0.992)	0.808	(0.00491)
	10%	0.156	(0.00379)	57.04	(1.22)	0.768	(0.00450)
RFV7	1%	0.22189	(0.0222)	4.569	(0.404)	0.978	(0.00122)
	2%	0.19436	(0.0125)	7.4709	(0.443)	0.967	(0.00123)
	5%	0.15661	(0.00601)	15.7523	(0.557)	0.947	(0.00136)
	10%	0.13121	(0.00342)	28.8085	(0.676)	0.927	(0.00143)
SCB7	2.5%	0.37075	(0.0319)	1.761	(0.117)	0.994	(0.000219)
	5%	0.30315	(0.0173)	2.581	(0.120)	0.990	(0.000271)
	10%	0.24114	(0.00853)	4.2302	(0.129)	0.984	(0.000310)
	25%	0.18275	(0.00346)	8.7751	(0.148)	0.972	(0.000341)

Table 6: Parameter estimates for logarithmic series distribution, ML-estimation, simulated means and standard errors, based on 1,000 simulations.

Set	π_s	$\hat{\phi}_s$	
RFV5	1%	0.778	(0.00580)
	2%	0.871	(0.00241)
	5%	0.942	(0.000755)
	10%	0.970	(0.000315)
RFV7	1%	0.579	(0.00978)
	2%	0.715	(0.00498)
	5%	0.847	(0.00186)
	10%	0.911	(0.000844)
SCB7	2.5%	0.293	(0.00913)
	5%	0.435	(0.00630)
	10%	0.584	(0.00374)
	25%	0.750	(0.00149)

Table 7: Parameter estimates for Pitman's sampling formula model, ML-estimation, simulated means and standard errors, based on 1,000 simulations.

Set	π_s	$\hat{\alpha}$		$\hat{\theta}$	
RFV5	1%	0.233	(0.0400)	496	(43.4)
	2%	0.167	(0.0260)	565	(32.1)
	5%	0.109	(0.0157)	639	(22.2)
	10%	0.0743	(0.0106)	690	(16.5)
RFV7	1%	0.440	(0.0370)	856	(88.4)
	2%	0.378	(0.0236)	1004	(65.5)
	5%	0.321	(0.0137)	1168	(44.0)
	10%	0.284	(0.00905)	1291	(33.2)
SCB7	2.5%	0.419	(0.0744)	5186	(837)
	5%	0.449	(0.0311)	4874	(384)
	10%	0.458	(0.0146)	4746	(200)
	25%	0.429	(0.00693)	5185	(112)

Table 8: Parameter estimates for zero- and right-truncated Poisson-lognormal model, ML-estimation,simulated means and standard errors, based on 1,000 simulations save SCB7 2.5% where 42 cases failed to converge properly. Right-trunction threshold as indicated.

Set	π_s	Trunc	$\hat{\mu}$		$\hat{\sigma}^2$		$\hat{\theta}$	
RFV5	1%	5	-2.22	(2.37)	4.5354	(5.89)	0.795	(0.361)
	2%	5	-2.44	(4.65)	8.298	(17.4)	0.789	(0.203)
	5%	10	-1.15	(2.51)	7.6426	(13.2)	0.797	(0.0458)
	10%	10	-0.632	(2.86)	10.2222	(30.2)	0.784	(0.0344)
RFV7	1%	5	-2.67	(0.626)	2.5288	(0.793)	0.951	(0.0172)
	2%	10	-2.18	(0.311)	2.7774	(0.420)	0.948	(0.00772)
	5%	10	-2.58	(0.480)	5.2128	(0.958)	0.912	(0.0164)
	10%	10	-2.60	(0.519)	7.0794	(1.34)	0.890	(0.0177)
SCB7	2.5%	5	-3.38	(0.434)	1.8981	(0.390)	0.975	(0.00658)
	5%	5	-3.98	(0.529)	3.2921	(0.593)	0.952	(0.0139)
	10%	5	-4.31	(0.757)	4.6501	(1.02)	0.923	(0.0267)
	25%	10	-3.34	(0.205)	4.5707	(0.302)	0.928	(0.00623)

Table 9: Parameter estimates for zero-truncated Poisson-inverse Gaussian model, ML-estimation, simulated means and standard errors, based on 1,000 simulations.

Set	π_s	$\hat{\mu}_s$		$\hat{\tau}_s$		$\hat{\theta}$	
RFV5	1%	0.770	(0.0815)	3.520	(0.240)	0.869	(0.0145)
	2%	1.54	(0.0857)	7.06	(0.333)	0.870	(0.00733)
	5%	3.58	(0.105)	19.1	(0.591)	0.860	(0.00411)
	10%	6.68	(0.124)	42.4	(0.956)	0.850	(0.00279)
RFV7	1%	0.119	(0.0486)	1.62	(0.112)	0.868	(0.143)
	2%	0.330	(0.0506)	3.04	(0.146)	0.931	(0.0114)
	5%	0.872	(0.0561)	7.45	(0.214)	0.936	(0.00419)
	10%	1.69	(0.0612)	15.2	(0.310)	0.934	(0.00241)
SCB7	2.5%	-	-	-	-	-	-
	5%	-	-	-	-	-	-
	10%	0.0636	(0.0230)	1.78	(0.0528)	0.841	(0.105)
	25%	0.242	(0.0207)	4.24	(0.0639)	0.914	(0.00748)

Table 10: Parameter estimates for zero-truncated Poisson-inverse Gaussian model, PF12 etimation, simulated means and standard errors, based on 1,000 simulations save RFV5 5%, RFV7 1% and SCB7 10% where 8, 8 and 36 cases respectively failed to converge.

Set	π_s	$\hat{\mu}_s$		$\hat{\tau}_s$		$\hat{\theta}$	
RFV5	1%	0.780	(0.109)	3.59	(1.07)	0.868	(0.0272)
	2%	1.57	(0.102)	10.3	(4.43)	0.859	(0.0177)
	5%	6.48	(3.53)	185	(429)	0.842	(0.0112)
	10%	-	-	-	-	-	-
RFV7	1%	0.190	(0.0679)	1.33	(0.218)	0.928	(0.0640)
	2%	0.365	(0.0740)	2.77	(0.441)	0.937	(0.0208)
	5%	0.860	(0.0720)	8.19	(1.50)	0.933	(0.00845)
	10%	1.78	(0.0906)	25.5	(7.00)	0.926	(0.00598)
SCB7	2.5%	-	-	-	-	-	-
	5%	-	-	-	-	-	-
	10%	0.0712	(0.0296)	1.75	(0.105)	0.850	(0.108)
	25%	0.323	(0.0297)	3.39	(0.200)	0.939	(0.00635)

B The PF12 Estimator Under the PiG Model

The estimation procedure amounts to finding the solution to the pair of equations

$$\begin{cases} \dfrac{t_1}{C - t_0} = \Pr\left(f_i = 1 \mid f_i > 0\right) \\ \dfrac{t_2}{C - t_0} = \Pr\left(f_i = 2 \mid f_i > 0\right) \end{cases}$$

From the recurrence formula of the PiG (see Carlson, 2002a, for details) we have

$$p_0 = \exp\left(\frac{\mu}{\tau}\left(1 - \eta\right)\right), \qquad p_1 = \frac{\mu}{\eta} p_0$$

and

$$p_2 = \frac{\mu\tau + \mu^2\eta}{2\eta^3} p_0.$$

Dividing the second equation by the first yields

$$\frac{t_2}{t_1} = \frac{p_2}{1 - p_0}\frac{1 - p_0}{p_1} = \frac{\tau + \mu\eta}{2\eta^2}$$

which after solving for μ yields

$$\mu = \frac{2t_2\eta^2 - t_1\tau}{t_1\eta}$$

and the equations can be re-written as

$$\begin{cases} h_1 = \dfrac{\tau}{\eta^2} + \dfrac{t_1}{C - t_0} \exp\left[\dfrac{2t_2\eta^2 - t_1\tau}{t_1\tau\eta}\left(\eta - 1\right)\right] - \left(\dfrac{2t_2}{t_1} + \dfrac{t_1}{C - t_0}\right) = 0 \\ h_2 = \mu - \dfrac{2t_2\eta^2 - t_1\tau}{t_1\eta} = 0 \end{cases}$$

Thus, we first find the value $\hat{\tau}$ that solves the first equation and then substitute this value for τ in the second to obtain an estimate for μ. Finding $\hat{\tau}$ requires numerical iteration methods.

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.

[2] Block, H and Olsson, L. (1976) Backwards Identification of Personal Information - Bakvägsidentifiering (in Swedish). *Statistisk Tidskrift*, **4**, pp. 135-144.

[3] Bulmer, M.G. (1974) On Fitting the Poisson Lnormal Distribution To Species-Abundance Data. *Biometrics*, **30**, pp. 101-110.

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

[5] Carlson, M. (2002b) On Assessing Disclosure Risk in Microdata. In J. Hagberg (Ed.) *Contributions to Social Network Analysis, Information Theory and Other Topics in Statistics, A Festschrift in Honour of Ove Frank.* pp. 202-213. Department of Statistics, Stockholm University.

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

[7] Cover, T.M. and Thomas, J.A. (1991) *Information Theory.* Monograph. New York: John Wiley.

[8] Dalenius, T. (1986) Finding a Needle In a Haystack or Identifying Anonymous Census Records. *Journal of Official Statistics*, **2**, pp. 329-336.

[9] 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.

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

[11] 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.

[12] Frank, O. (1983) Statistical Disclosure Control. *Statistical Review*, **5**, *Essays in Honour of Tore E. Dalenius*, pp. 173-178.

[13] 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 Intern. Fed. of Classification Soc. (IFCS), Tech. Univ. Aachen, F.R.G., 29 June - 1 July 1987,* pp. *687-690.* New York: North-Holland.

[14] Greenberg, B.V., Zayatz, L.V. (1992) Strategies for Measuring Risk in Public Use Microdata Files. *Statistica Neerlandica*, **46**, pp. 33-48.

[15] 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.

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

[17] Hoshino, N. and Takemura, A. (1998) On the Relation Between Logarithmic Series Model and Other Superpopulation Models Useful for Microdata Disclosure Risk Assessment. Discussion Paper, 98-F-7, Faculty of Economics, University of Tokyo. (Published in *Journal of Japan Statistical Society*, **28**, pp. 125-134, 1998, after revision)

[18] Johnson, N.L., Kotz, S. and Kemp, A,W. (1992) *Univariate Discrete Distributions*, 2nd ed.. Monograph. New York: John Wiley & Sons.

[19] Paass, G. (1988) Disclosure Risk and Disclosure Avoidance for Microdata. *Journal of Business & Economic Statistics*, **6**, pp. 487-500.

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

[21] 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.

[22] Skinner, C.J. and Elliot, M.J. (2002) A Measure of Disclosure Risk in Microdata. *Journal of the Royal Statistical Society, Series B*, **64**, pp. 855-867.

[23] Skinner, C.J. and Holmes, D.J. (1993) Modelling Population Uniqueness. In *Proceedings of the International Seminar on Statistical Confidentiality, Dublin, 8-10 September1992*, pp. 175-199. Luxembourg: Office for Official Publications of the European Communities.

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

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

[26] St-Cyr, P. (1998) Modelling Population Uniqueness Using a Mixture of Two Distributions. In *Statistical Data Protection - Proceedings of the Conference, Lisbon, 25 to 27 March 1998 - 1999 edition*, pp. 277- 286. Luxembourg: Office for Official Publications of the European Communities.

[27] 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.