Research Report

Department of Statistics

No. 2002:7

Assessing Microdata Disclosure Risk Using the Poisson-Inverse Gaussian Distribution

Michael Carlson

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

Assessing Microdata Disclosure Risk Using the Poisson-Inverse Gaussian Distribution

Michael Carlson[*]

December 18, 2002

Abstract

An important measure of identification risk associated with the release of microdata or large complex tables is the number or proportion of population units that can be uniquely identified by some set of characterizing attributes which partition the population into subpopulations or cells. Various methods for estimating this quantity based on sample data have been proposed in the literature by means of superpopulation models. In the present paper the Poisson-inverse Gaussian (PiG) distribution is proposed as a possible approach within this context. Disclosure risk measures are discussed and derived under the proposed model as are various methods of estimation. An example on real data is given and the results indicate that the PiG model may be a useful alternative to other models.

Keywords: statistical disclosure; uniqueness; inverse-Gaussian; Poisson-mixture; superpopulation.

1 Introduction

A considerable amount of research has been done in the area of statistical disclosure and different approaches to defining and assessing disclosure risk are treated in depth by among others Dalenius (1977), Duncan and Pearson (1991), 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). A special case concerns the release of public-use microdata files and so-called identity disclosure, see Duncan and Lambert

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

(1989). 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, see e.g. Block and Olsson (1976) and Dalenius (1986), although Blien et al. (1993) demonstrated that it may be difficult in practice. The main concern is to ensure that no record in a released microdata set can be reliably associated with an identifiable individual.

For instance, a unique is defined as an entity that has a unique set of values on a set of characterizing attributes or key attributes. A unit that is unique in the population is referred to as a population unique whereas a unit that is unique in a 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. Obviously a population unique is subjected to a greater risk of being exposed relative to other non-unique units if included in the released data. Furthermore, 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, see e.g. Skinner et al. (1994) and Elliot et al. (1998). Thus, a possible indicator of the identification risk associated with the release of microdata is the number or proportion of population uniques included in the sample amongst the sample uniques.

The objective is to estimate this proportion based on sample information, e.g. a data set considered for release. Various methods for estimating this quantity based on sample data have been proposed in the literature by means of superpopulation models and especially compound Poisson models. 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. Bethlehem et al. (1990) were perhaps the first to adapt a superpopulation approach and others include Chen and Keller-McNulty (1998), Hoshino (2001), Samuels (1998), Skinner and Holmes (1993, 1998), St-Cyr (1998) and Takemura (1999).

In the present paper we propose the Poisson-inverse Gaussian (PiG) distribution as a possible candidate. This distribution has appeared elsewhere in the literature but we are not aware of it being applied to the disclosure problem earlier. It was introduced by Holla (1966) in studies of repeated accidents and recurrent disease symptoms. Sichel (1971) developed the PiG to a more general three-parameter family of distributions and applied it to density and size distributions of diamonds, sentence-length and word frequency data and to model repeat-buying behavior, (Sichel, 1973, 1974, 1975, and 1982a). Ord and Whitmore (1986) evaluated the PiG as an alternative to other distributions for species abundance data and Willmot (1987) for modelling insurance claim data. Chen and Keller-McNulty (1998) noted that, in practice, the frequency distribution in disclosure applications tends to have an inverse J-shape with heavy upper tail. St-Cyr (1998) also describes this typical behavior. 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. Furthermore, the PiG distribution is expressed in closed form which gives it an advantage over e.g. the lognormal which requires numerical integration.

Here we will limit the scope to a theoretical discussion of the PiG model with a simple example on real data to illustrate the method. An evaluation of the model with real-life data examples and its competitiveness with alternative approaches is intended to appear in a separate report. In the following section some basic notation is introduced and the superpopulation model is specified. In section 3 the PiG distribution is reviewed and in section 4 its application to the problem of assessing disclosure risks, here in terms of uniqueness, is discussed. Parameter estimation is described in section 5 and in section 6 the results of the empirical example are reported. Some concluding remarks and directions for future research are given in section 7.

2 Specification of the Superpopulation Model

2.1 Basic 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 $\pi_s = n/N$. With each unit $h \in U$ is associated the values of a number of discrete variables, $Z_1, \ldots, Z_q$ with $C_1, \ldots, C_q$ categories respectively. The cross-classification of these variables define the discrete variable X with $\Pi C_i = C$ categories or cells and for simplicity we let the cells of X be labeled as $1, 2, \ldots, C$.

Following e.g. Bethlehem et al. (1990) the Z_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. The sample counterpart is analogously defined and denoted by f_i.

Define T_j and its sample counterpart t_j as the number of cells of size j, i.e.

$$T_j = \sum_{i=1}^{C} I_{F_i=j} = \#\left(i; F_i = j\right), \qquad j = 0, 1, \ldots, N$$

and

$$t_j = \sum_{i=1}^{C} I_{f_i=j} = \#\left(i; f_i = j\right) \qquad j = 0, 1, \ldots, n$$

respectively and where $I_{(\cdot)}$ denotes the usual indicator function. 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). It is clear that

$$\sum_{i=1}^{C} F_i = \sum_{j=1}^{N} jT_j = N, \qquad \sum_{i=1}^{C} f_i = \sum_{j=1}^{n} jt_j = n$$

and

$$\sum_{j=0}^{N} T_j = \sum_{j=0}^{n} t_j = C.$$

Of these quantities, C, N and n are fixed in the design, the f_i and t_j are observed and the F_i and T_j are assumed to be unknown. The goal is to model and estimate the population frequency structure, i.e. the T_j and especially T_1 which is the number of unique individuals in the population, based on sample information.

2.2 Superpopulation model

The frequency structure, $\{T_j\}$, is a function of the actual F_i which are unknown and therefore need to be estimated from the observed f_i and t_j. However, as St-Cyr (1998) commented, it is not possible for a sample to carry all the information about the structure of a population and finite population theory will give only unreliable estimates when the sampling fraction is small. See also the discussion by Willenborg and de Waal (2000, p. 55). A way out is to model the frequency structure and view the population of cell frequencies as a realization of a superpopulation model.

As a starting point we therefore assume that the cell frequencies are 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 we view the λ_i 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 simplification seems reasonable also in view of that we are not conditioning on any of the characterizing variables defining the key.

The specification of the mixing distribution $g(\lambda)$ is the crucial step and several different suggestions have been studied. Bethlehem et al. (1990) proposed a gamma distribution which implies that the marginal distribution of each F_i is the negative-binomial. This model has however been noted to provide a poor fit to real-life data. Skinner and Holmes (1993, 1998) argue instead for the use of a lognormal distribution. Chen and Keller-McNulty (1998) considered a shifted negative-binomial for the marginal distribution of the F_i and achieve better results compared to the Poisson-gamma. St-Cyr (1998) proposed a mixture of a Pareto and a truncated lognormal distribution based on his findings from studying the relationship between the conditional probability of population uniqueness and the sampling fraction. Hoshino and Takemura (1998) investigated the relationships between different models and provide interesting results.

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. Skinner and Holmes (1993) specified their superpopulation model to allow for individual rates to be zero with positive probability. Following their idea we therefore assume that the distribution of the λ_i is a mixture θ of a discrete probability mass at zero and $(1-\theta)$ of a continuous distribution with pdf $g(\lambda)$ on $(0, \infty)$, i.e. we specify the marginal of the cell frequencies as

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

where $0 \leqslant \theta < 1$ and

$$P_j = \int_0^\infty \frac{\lambda^j e^{-\lambda}}{j!} g(\lambda)\, d\lambda. \tag{2}$$

Note that with the Poisson assumption the total of the cell counts ΣF_i is a random variable. Although it is true from the design that $\Sigma F_i = N$ it may suffice to check if $\Sigma E(F_i) = N$ rather than conditioning on the actual population size, cf. Bethlehem et al. (1990). Denoting the expectation of the distribution in (2) by μ this requirement translates to

$$C(1-\theta)\mu = N. \tag{3}$$

3 The Poisson-Inverse Gaussian Distribution

We assume that the individual rates follow an inverse Gaussian (iG) distribution and using the same parameterization as Willmot (1987) the pdf is

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

where $\mu > 0$, $\tau > 0$. The mean and variance of the IG are μ and $\mu\tau$ respectively. Folks and Chhikara (1978) provide a review of the iG distribution with an extensive set of references. See also Johnson et al. (1994, chapter 15). The inverse Gaussian distribution appears e.g. as the first passage time distribution of Brownian motion with positive drift. It may also be derived through an inversion relationship associated with cumulant generating functions of the Gaussian and inverse Gaussian families, hence the name.

Integrating out λ from (2) with $g(\lambda)$ replaced by (4) yields the probability mass function (pmf) of the Poisson-inverse Gaussian (PiG), i.e.

$$P_j = \frac{\sqrt{\eta}}{j!}\left(\frac{\mu}{\eta}\right)^j \frac{K_{j-1/2}(\mu\eta/\tau)}{K_{-1/2}(\mu/\tau)}, \quad j = 0, 1, \ldots \tag{5}$$

where $\eta = \sqrt{1+2\tau}$ and $K_\gamma(z)$ denotes a modified Bessel function of the third kind (sometimes referred to as the second kind) of order γ and with argument z, see Abramowitz and Stegun, (1970, chapter 10). The mean and variance of the PiG distribution are μ and $\mu(1+\tau)$ respectively.

The expression in (5) is not always practical due to the Bessel functions, but by using that $K_{-1/2}(z) = K_{1/2}(z) = \sqrt{\pi/2z}\exp(-z)$ we note that the first two probabilities are

$$P_0 = \exp\left(\frac{\mu}{\tau}(1-\eta)\right) \quad \text{and} \quad P_1 = \frac{\mu}{\eta}P_0. \tag{6}$$

Using also that $K_{-\gamma}(z) = K_{\gamma}(z)$ and the recurrence relationship $K_{\gamma+1}(z) = (2\gamma/z) K_{\gamma}(z) + K_{\gamma-1}(z)$, a more practical recurrence formula for calculating the probabilities is given by

$$P_j = \frac{\tau}{\eta^2} \frac{2j-3}{j} P_{j-1} + \frac{\mu^2}{\eta^2} \frac{1}{j(j-1)} P_{j-2}, \quad j = 2, 3, \ldots \tag{7}$$

3.1 Sampling From the Population

Assume that the population level cell frequencies are generated from a PiG model. Under simple random sampling without replacement, the sampling distribution of the cell frequencies of the PiG is hard to manipulate. We therefore assume Bernoulli sampling, cf. e.g. Särndal et al. (1992, chapter 3), in which each unit is drawn independently from the population with equal probability $\pi_s = n/N$ as a convenient approximation. This yields

$$f_i \mid \lambda_i \sim Po(\pi_s \lambda_i) \quad \text{and} \quad F_i - f_i \mid \lambda_i \sim Po((1 - \pi_s) \lambda_i)$$

and

$$f_i \mid F_i \sim Bin(F_i, \pi_s). \tag{8}$$

It is then easily seen that the marginal pmf for the sample cell frequencies f_i is defined by

$$\Pr(f_i = j) = \theta I_{j=0} + (1-\theta) p_j \tag{9}$$

where

$$\begin{aligned} p_j &= \int_0^\infty \frac{(\pi_s \lambda)^j e^{-\pi_s \lambda}}{j!} \frac{\mu}{(2\pi\tau\lambda^3)^{1/2}} \exp\left(-\frac{(\lambda-\mu)^2}{2\tau\lambda}\right) d\lambda \\ &= \int_0^\infty \frac{\lambda^j e^{-\lambda}}{j!} \frac{\mu_s}{(2\pi\tau_s\lambda^3)^{1/2}} \exp\left(-\frac{(\lambda-\mu_s)^2}{2\tau_s\lambda}\right) d\lambda \end{aligned} \tag{10}$$

i.e. a PiG distribution where $\mu_s = \pi_s \mu$, $\tau_s = \pi_s \tau$ and defining $\eta_s = \sqrt{1 + 2\tau_s}$; the second line in (10) is derived by simple variable substitution. This provides an easy transformation when we have a sample from the larger population, i.e. given the sample we estimate the parameters μ_s and τ_s and simply multiply by π_s^{-1}. See also section 2 of Sichel (1982a) and the discussion concerning sampling in Takemura (1999).

4 Risk assessment

The outline of the disclosure problem considered here is the same as that of many authors, e.g. Bethlehem et al. (1990), Elliot et al. (1998), Paass (1988) 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. 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_i is the number of units belonging to subpopulation U_i and we let $i\left(r\right)$ denote the value of X for record r. If $F_{i(r)}$ was known the intruder could infer that the probability of a correct link is $F_{i(r)}^{-1}$ and if $F_{i(r)} = 1$ the link is correct with absolute certainty.

Usually the intruder will not know the true value of $F_{i(r)}$ since the microdata set contains only a sample but by introducing a superpopulation model he may attach a probability distribution $Pr\left(F_i = j\right)$ to the cell frequencies. Furthermore, 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. So equating disclosure risk with uniqueness, a simple measure of the risk for a given sample is the proportion of sample uniques that are also population uniques, i.e.

$$R = \frac{\#\left(\text{records that are Pop.Uniques } \textit{and} \text{ SampleUniques}\right)}{\#\left(\text{records that are SampleUniques}\right)}. \tag{11}$$

Under simple random sampling or Bernoulli sampling the expected number of population uniques to fall into the sample is $\pi_s T_1 = nT_1/N$. Since T_1 is assumed unknown, the ratio in (11) will have to be estimated. Under the model (1) we have that the expected number of population uniques is

$$E\left(T_1\right) = C\Pr\left(F_i = 1\right) = C\left(1-\theta\right)P_1.$$

Thus, an obvious risk measure denoted by R_1, is defined as the proportion of the observed number of sample uniques expected to be population uniques, i.e.

$$R_1 = \frac{E\left(T_1\right)/N}{t_1/n} = \frac{\pi_s C\left(1-\theta\right)}{t_1}\frac{\mu}{\eta}\exp\left(\frac{\mu}{\tau}\left(1-\eta\right)\right) \tag{12}$$

where we have used (6). Again assuming Bernoulli sampling, an alternative risk measure R_2 follows naturally from the conditional pmf of F_i given f_i, i.e.

from (5), (8) and (10) we have

$$R_2 = \Pr(F_i = 1 \mid f_i = 1) = \frac{\pi_s \Pr(F_i = 1)}{\Pr(f_i = 1)} = \frac{E(T_1)/N}{E(t_1)/n} \tag{13}$$

which simplifies to the risk measure

$$R_2 = \frac{\eta_s}{\eta} \exp\left(\frac{\mu}{\tau}(\eta_s - \eta)\right) \tag{14}$$

i.e. the observed value of t_1 in (12) is replaced for its expectation under the model. This is the approach discussed by Skinner and Holmes (1998). Note that $R_2 \to 1$ as $\pi_s \to 1$ which is not necessarily the case with R_1. The risk measures R_1 and R_2 coincide if and only if a perfect fit of the model to the observed number of sample uniques is obtained, i.e. $E(t_1) = t_1$. Either choice, the disclosure risk is estimated by replacing the parameters by their estimates into (12) or (14).

4.1 Extended Risk Measures

It should be noted that population uniqueness is neither a sufficient nor necessary condition for re-identification or for disclosing additional information, see e.g. Frank (1976) and Willenborg and de Waal (1996, pp. 19-20). It is not a sufficient condition since, first of all, the unique unit must be included in the sample and secondly, it must also be known to the intruder that the unit is in fact unique. It is not necessary for several reasons. If for instance a person in the population shares the same values on the key attributes with say only one other person, they will both be able to re-identify and disclose information about each other. In general, coalitions of respondents exchanging information can be formed within small groups sharing the same scores on the key attributes, in order to disclose information about an individual within the same group but outside the coalition. Alternatively, if a group of people share the same values on the key attributes, none of them are unique. But if they in addition all share the same score on a certain sensitive attribute provided in the released data, the sensitive information can be disclosed for all the individuals in that group without re-identifying individual records. Another possibility is response knowledge, i.e. knowledge that a specific individual participated in the survey and consequently that his or her data must be included in the data. Identification and disclosure can then occur if the person is unique in the sample and not necessarily in the population (Bethlehem et al., 1990).

These issues will not be investigated further in the present paper, but they do however motivate an extension of the risk measure in (13) to a more general measure defined by

$$\Pr\left(F_i = j + k \mid f_i = j\right) = \frac{1}{k!} \frac{\int_0^\infty \left(1 - \pi_s\right)^k \lambda^{k+j} \exp\left(-\lambda\right) g\left(\lambda\right) d\lambda}{\int_0^\infty \lambda^j \exp\left(-\pi_s \lambda\right) g\left(\lambda\right) d\lambda}$$

for $j = 1, 2, \ldots$ and $k = 0, 1, \ldots$. Simple examples pertaining to some of the situations just described may be e.g. $j = 1$ and $k = 1$ which after simplifying yields

$$\Pr\left(F_i = 2 \mid f_i = 1\right) = \left(1 - \pi_s\right) \frac{\left(\mu\eta + \tau\right)}{\eta^2} R_2$$

or $j = 2$ and $k = 0$ which yields

$$\Pr\left(F_i = 2 \mid f_i = 2\right) = \pi_s \frac{\eta_s^2 \left(\mu\eta + \tau\right)}{\eta^2 \left(\mu_s \eta_s + \tau_s\right)} R_2$$

where R_2 is the risk measure defined in (14).

5 Estimation

5.1 Moment Based Estimators

Let S_0 denote the number of structural zeroes. If this number is known a priori the parameter θ is known and equals S_0/C. When this is the case and especially when $\theta = 0$, the parameters μ_s and τ_s can be estimated using simple moment approaches. Simple moment estimators are given by the sample mean and variance, i.e.

$$\tilde{\mu}_s = \frac{n}{C - S_0} \tag{15}$$

and

$$\tilde{\tau}_{s,1} = \frac{1}{\tilde{\mu}_s \left(C - S_0\right)} \sum_{j=0}^{\infty} j^2 t_j - \tilde{\mu}_s - 1$$

with t_0 replaced for $t_0 - S_0$. The latter was however shown not to be very efficient (cf. Sichel 1982b) and a more efficient and equally simple estimate is obtained by matching the mean and the proportion of empty cells to those of the underlying distribution, yielding

$$\tilde{\tau}_{s,2} = \frac{2\tilde{\mu}_s \left(\tilde{\mu}_s + \log\left(\frac{t_0}{C - S_0}\right)\right)}{\left(\log\left(\frac{t_0}{C - S_0}\right)\right)^2}. \tag{16}$$

In practice it would however more often be the case that S_0 is unknown and that θ needs to be considered in the estimation process. By employing a zero-truncated approach where only the non-empty sample cell frequencies are considered, this problem is circumvented and θ is accordingly treated as a nuisance parameter since

$$\Pr\left(f_i = j \mid f_i \geq 1\right) = \frac{(1-\theta)\, p_j}{1 - (\theta + (1-\theta)\, p_0)} = \frac{p_j}{1-p_0}, \quad j = 1, 2, \ldots \tag{17}$$

Zero-truncated estimation was described by Sichel (1975, 1982b) who obtained an efficient estimator by matching the average cell size and the proportion of uniques, both amongst the non-empty cells, to the underlying distribution. The estimation procedure entails solving the equation

$$(1+g)\ln g - Ag + B = 0 \tag{18}$$

for g and where

$$A = \frac{2n}{(C-t_0)} - \ln\frac{n}{t_1} \qquad \text{and} \qquad B = \frac{2t_1}{(C-t_0)} + \ln\frac{n}{t_1}.$$

Equation (18) is easily solved by numerical iteration, e.g. Newton-Raphson, and from the solution $\tilde{g}$ we obtain estimates of μ and τ as

$$\tilde{\mu}_{s,ztr} = \frac{1+\tilde{g}}{2\tilde{g}} \ln\left(\frac{\tilde{g}n}{t_1}\right) \tag{19}$$

and

$$\tilde{\tau}_{s,ztr} = \frac{1-\tilde{g}^2}{2\tilde{g}^2} \tag{20}$$

respectively. An initial estimate to start the iteration is given by the estimates of τ in (16) and then using (20).

5.2 Maximum Likelihood

Maximum likelihood (ML) estimation is fairly straightfoward for the PiG model. When the number of structural zeroes is known a priori, the likelihood is derived from (10) over the $C - S_0$ non-structural-zero cells. Willmot (1987) gave the ML-estimates for the present parameterization and we include them here for the sake of self-containment. The loglikelihood is

$$l_{ML} = \sum_{j=0}^{\infty} t_j \log p_j$$

with t_0 replaced for $t_0 - S_0$ and it is easily shown that the ML-estimate of μ_s is simply the average cell size,

$$\hat{\mu}_s = \frac{n}{C - S_0}. \tag{21}$$

The ML-estimate of τ_s is the solution to

$$h = \sum_{j=0}^{\infty} t_j \varphi_j - n = 0 \tag{22}$$

with t_0 replaced for $t_0 - S_0$ and where

$$\varphi_j = \frac{(j+1)\, p_{j+1}}{p_j}.$$

The values of φ_j are conveniently computed from the following recursions which are a direct consequence of (6) and (7):

$$\varphi_0 = \frac{\mu_s}{\eta_s} \quad \text{and} \quad \varphi_j = \left(\frac{\tau_s\,(2j-1)}{\mu_s^2} + \frac{1}{\varphi_{j-1}} \right) \varphi_0^2, \quad j = 1, 2, \ldots$$

Equation (22) is easily solved using e.g. Newton-Raphson iteration and the required derivative of h with respect to τ_s is

$$\frac{\partial h}{\partial \tau_s} = \frac{1+\tau_s}{\tau_s^2} \sum_{j=0}^{\infty} t_j \varphi_j \left(\varphi_{j+1} - \varphi_j \right) - \frac{1}{\tau_s} \sum_{j=0}^{\infty} t_j \varphi_j$$

with μ_s replaced by (21). An initial estimate to start the iteration is given by the estimator in (16).

As mentioned in the preceding subsection it would more often be the case that θ is unknown and needs to be considered in the estimation process. By employing a zero-truncated likelihood where only the non-empty sample cell frequencies are considered, θ is treated as a nuisance parameter as shown in (17). This is the approach considered by Skinner and Holmes (1993) designating it a conditional likelihood (CL). The loglikelihood of the f_i for those i which $f_i \geq 1$, is thus defined as

$$l_{CL} = \sum_{j=1}^{\infty} t_j \log \frac{p_j}{1 - p_0} = \sum_{j=1}^{\infty} t_j \log p_j - (C - t_0) \log (1 - p_0) \tag{23}$$

which yields the system of equations

$$\begin{cases} h_1 = \dfrac{\mu_s}{1 - p_0} - \dfrac{n}{C - t_0} = 0 \\ h_2 = \displaystyle\sum_{j=1}^{\infty} t_j \varphi_j - n + \frac{n p_0}{\eta_s} = 0 \end{cases} \tag{24}$$

Estimates of μ_s and τ_s are the solutions to (24) and may be obtained by numerical iteration methods such as Newton-Raphson. The derivation of (24) and the required derivatives are provided in the appendix. Small scale experiments indicate however that the rate of convergence for the zero-truncated approach may be slow and that some improvement of the numerical method used may be called for. In our experience using (19) and (20) as initial estimates will usually be a good choice.

5.3 Right-truncation

Skinner and Holmes (1993) also considered truncating the set of probabilities in (17) above a threshold value m. The idea is 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. A further motivation for this approach is a possible reduction of computational effort. Thus, the p_j are assumed to be proportional to (10) for $j = 1, \ldots, m$ and no assumptions are made about the p_j for $j \geqslant m + 1$. Define

$$q_m^* = \Pr\left(f_i > m \mid f_i \geqslant 1\right) = \frac{1 - \sum_{j=0}^{m} p_j}{1 - p_0}.$$

The right-truncated version of (23) is then expressed as

$$l_{trCL} = \sum_{j=1}^{m} t_j \log\left(\frac{p_j/\left(1-p_0\right)}{1-q_m^*}\right) = \sum_{j=1}^{m} t_j \log p_j - t_m^* \log p_m^*$$

yielding the system of equations

$$\left\{\begin{array}{l} h_1^* = \sum_{j=1}^{m} \dfrac{jp_j}{p_m^*} - \dfrac{n_m^*}{t_m^*} = 0 \\ h_2^* = \sum_{j=1}^{m} t_j \varphi_j - n_m^* - \dfrac{t_m^*}{p_m^*}\left(m+1\right) p_{m+1} + \dfrac{t_m^*}{p_m^*} p_1 = 0 \end{array}\right. \tag{25}$$

where

$$t_m^* = \sum_{j=1}^{m} t_j, \quad n_m^* = \sum_{j=1}^{m} j t_j \quad \text{and} \quad p_m^* = \sum_{j=1}^{m} p_j.$$

Finding the solutions to (25) requires numerical iteration. E.g. Newton-Raphson requires the derivatives of (25) and it is straightfoward to derive these using the results in the appendix but the result is however not very elegant and convergence may be slow. A further problem indicated by small scale experiments on simulated data seems to be that the truncated approach is sensitive to the choice of starting values in combination with the selected threshold value; depending on the choice the iteration may or may not converge.

As an alternative one might consider the method proposed by Chen and Keller-McNulty (1998) who fit their model to the observed values of t_1 and t_2. Estimators based on their idea, which we will denote by PF12, are thus defined as the solutions to the system of equations

$$\begin{cases} \dfrac{p_1}{1-p_0} = \dfrac{t_1}{C-t_0} \\ \dfrac{p_2}{1-p_0} = \dfrac{t_2}{C-t_0} \end{cases} \tag{26}$$

and is motivated by the same line of reasoning motivating the right-truncation approach. Finding the solutions requires numerical iteration methods such as Newton-Raphson iteration and the required derivatives are straightfoward to derive using the results in the appendix.

5.4 Estimation of θ

The zero-truncated approaches imply estimation of θ. From (9) we have $E(t_0) = C\theta + C(1-\theta)p_0$ so once the estimates of μ_s and τ_s are obtained it is straightfoward to estimate θ by replacing $E(t_0)$ for t_0 and p_0 for its estimate, i.e.

$$\hat{\theta} = \frac{t_0 - C\hat{p}_0}{C(1-\hat{p}_0)}.$$

As a consequence we have $\hat{t}_0 = t_0$, that is, we obtain a perfect fit for the number of empty cells in the sample. Furthermore the restriction in (3) is automatically satisfied. For example using (24), the zero-truncated likelihood method yields $\hat{\mu}_s = n(1-\hat{p}_0)/(C-t_0)$, and remembering that $\hat{\mu} = N\hat{\mu}_s/n$, it is seen that

$$C\left(1-\hat{\theta}\right)\hat{\mu} = C\left(1 - \frac{t_0 - C\hat{p}_0}{C(1-\hat{p}_0)}\right)\frac{N(1-\hat{p}_0)}{C-t_0} = N.$$

In applications one should be aware of the possibility of obtaining negative estimates of θ which occurs if $t_0 < C\hat{p}_0$. This implies that the number of structural zeroes is negative and indicates that the estimation procedure is over-adjusting to the data. In such cases or if $\hat{\theta}$ is close to zero one may assume that $\theta = 0$ and use the ordinary ML procedure as described above.

6 An Empirical Example

6.1 Description of the Data

The data was provided by Statistics Sweden and originates from the 1990 census in Sweden. It consisted of frequency distributions, T_0, T_1, $T_2, \ldots$ for $N = 160,536$ individuals of ages 20-65 residing in Uppsala county. The following key variables were used: municipal (6), sex (2), age in one-year bands (46), marital status (10), citizenship, Swedish or foreign, (2) and income in 10,000 SEK bands (176). The numbers in parenthesis indicate the number of observed categories of the respective variables from which the total number of combinations is given as $C = 1,943,040$. Of these $T_0 = 1,903,218$ were found to be empty leaving a total of $39,822$ observed combinations. The number of population uniques T_1 was $19,273$ and the largest cell contained 140 units (one cell). Figure 1 displays the population cell size distribution save T_0 and we note the inverse J-shape and the heavy right tail. To illustrate the inverse J-shape in more detail the first 30 cell sizes are also shown.

From the data set a simple random sample without replacement of size $n = 16,054$ was drawn ($\pi_s = 0.1$). The largest cell size of the $10,046$ non-empty cells in the sample was 18 (one cell). The observed number of sample uniques was $t_1 = 7,216$ or approximately 45% of the sample. Of these $1,952$ where found to be true population uniques which is the expected number given the sampling fraction, i.e. approximately 10% of T_1. Thus the ratio (11) defining the risk measure is 0.2705 which is the quantity to be estimated. The sample cell frequencies are provided in table 2.

6.2 Results

Four variations of the PiG-model and methods of estimation were fitted to the data. Given the large number of empty cells it may seem natural to consider only models that truncate at zero but for sake of illustration both the ordinary and the zero-truncated PiG-models were fitted. In both cases we used ML estimation using Newton-Raphson iteration and the moment based estimators (15-16) and (19-20) respectively as starting values. The PF12 estimation procedure for the zero-truncated case was included in the study as well. The equations in (26) were solved by Newton raphson iteration using (19-20) as starting values. We also experimented with the right-truncation method as described above trying various threshold values. Initially we used Newton-Rapson iteration but it was found that better and faster results for this data set were obtained by using the Matlab (2001) minimization routine `fminsearch` which builds on a simplex direct search method. The

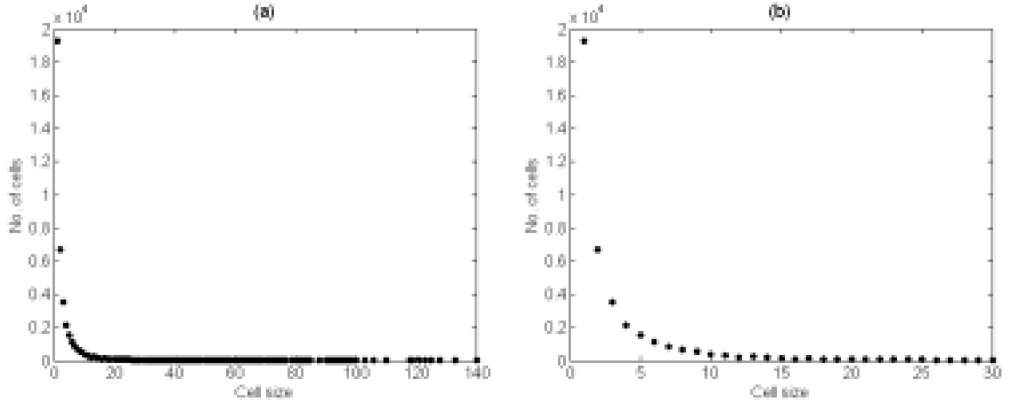

Figure 1: Population cell size distributions of the example data set. Display (a) shows the entire distribution and (b) shows the details of the left-hand-side. The frequency of empty cells, i.e. T_0, is omitted in both cases.

routine may however result in local minima so some experimentation with starting values may be required. Here we used the estimators (19-20) without encountering any problems. Choosing the threshold $m = 5$ was found to yield best results in terms of goodness-of-fit measures and when comparing the estimated risk ratio to the true risk ratio.

For comparison two alternative models were considered for the data. The first was Fisher's logarithmic series distribution (LSD). Taking the mixing distribution $g(\lambda)$ in (2) to be a gamma distribution results in the negative-binomial (NB) distribution. In many cases it has however been noted that the α parameter of the NB tends to be very small in disclosure applications. In such cases it may be appropriate to consider instead the limiting distribution of the zero-truncated NB as $\alpha \to 0$ which results to the LSD. The pdf of the LSD is defined by

$$\Pr(F_i = j) = -\frac{\phi^j}{j \log(1-\phi)}, \qquad j = 1, 2, \ldots$$

where $0 < \phi < 1$. Assuming Bernoulli sampling the marginal distribution of the cell sizes is also LSD with parameter

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

It can be shown that R_2 is simplified to

$$R_2 = -\frac{n}{C - t_0} \frac{(1-\phi)\log(1-\phi_s)}{\phi_s}.$$

Table 1: Estimates of model parameters, number of population uniques T_1, and risk measure R_2 and loglikelihood.

Model	Parameters			Loglikel.	$\hat{\mathrm{T}}_1$	$\hat{\mathrm{R}}_2$
Logarith. series dist., ML	$\hat{\phi}_s = 0.583$			-5169.0	10724	0.1601
Poisson-lognormal	$\hat{\mu}_s =$	$\hat{\sigma}^2 =$	$\hat{\theta} =$			
(1) z-tr, cens. m = 4, ML	-3.331	3.247	0.951	-9253.7	16646	0.2306
(2) z-tr, r-tr, m = 5, ML	-3.622	3.657	0.945	-8206.2	17366	0.2419
Poisson-inverse Gaussian	$\hat{\mu}_s =$	$\hat{\tau}_s =$	$\hat{\theta} =$			
(1) ML	0.008	1.893	–	-72972.4	25286	0.3448
(2) z-tr, ML	0.074	1.750	0.889	-10058.7	21636	0.2999
(3) z-tr, PF12	0.117	1.552	0.931	-10062.4	19629	0.2720
(4) z-tr, r-tr, m = 5, ML	0.106	1.476	0.924	-8207.9	20348	0.2793

The LSD model was fitted using ordinary ML estimation and Newton-Raphson iteration.

The second alternative model was the zero-truncated Poisson-lognormal (PLN) distribution proposed by Skinner and Holmes (1993, 1998). The model is defined by choosing $g(\lambda)$ in (2) to be distributed as lognormal with parameters μ and $\sigma^2 < 0$. Assuming Bernoulli sampling the marginal distribution of the cell sizes is also PLN with parameters $\mu_s = \mu + \log \pi_s$ and $\sigma_s^2 = \sigma^2$. Unfortunately the PLN distribution is not available in closed form so numeric integration is required to calculate the probabilities and the risk measure R_2. For this data set 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. Skinner and Holmes suggested either censoring or truncating the loglikelihood above a threshold value m. We tried both methods on the sample data and found that choosing $m = 4$ for the censored version and $m = 5$ for the right-truncated version yielded best results in terms of goodness-of-fit measures and in comparison to the true risk ratio. Maximizing the censored and truncated loglikelihoods also required some experimentation including Newton-Raphson and the Nelder-Mead method mentioned above. Both methods were found to be sensitive to the choice of starting values and the latter occasionally produced negative estimates of σ^2.

To compare the fit to the different models two conventional goodness-of-fit

Table 2: Observed and fitted cell size frequencies and goodness-of-fit statistics for sample data set. The (*) indicates collapsing of categories above and including the corresponding cell size. The models are LSD: logarithmic series distribution and ML estimation, PLN: (1) zero-truncated and censored likelihood, $m = 4$, (2) zero- and right-truncated likelihood, $m = 5$, PiG: (1), full likelihood, $\theta = 0$, (2), zero-truncated likelihood, (3), zero-truncated and the PF12 estimator, (4), zero- and right-truncated likelihood, $m = 5$.

		Fitted $\hat{t}_j$						
Size	Observ.	LSD	PLN		PiG			
j	t_j		(1)	(2)	(1)	(2)	(3)	(4)
0	1932994	–	–	–	1932993.2	–	–	–
1	7216	6697.2	7217.7	7220.3	7300.8	7216.5	7216	7218.3
2	1573	1951.7	1561.4	1550.4	1457.6	1529.5	1573	1540.0
3	533	758.3	555.7	562.7	576.5	596.3	598.8	578.6
4	272	331.5	258.0	267.2	285.0	290.0	283.5	270.5
5	155	154.6	453.2*	148.4	157.8	157.9	150.2	141.5
6	117	75.1	–	–	93.6	92.1	85.2	–
7	70	37.5	–	–	58.2	56.3	50.6	–
8	41	19.1	–	–	37.4	35.6	31.1	–
9	36	9.9	–	–	24.7	23.1	19.6	–
10	11	5.2	–	–	16.6	15.2	12.6	–
11	8	2.8	–	–	11.3	10.2	8.2	–
12	4	1.5	–	–	7.8	7.0	5.5	–
13	5	1.7*	–	–	5.5	4.8	3.6	–
14	3	–	–	–	3.9	3.3	2.5	–
15	1	–	–	–	2.8	2.3	1.7	–
16	0	–	–	–	2.0*	5.8*	3.8*	–
17	0	–	–	–	–	–	–	–
18	1	–	–	–	–	–	–	–
19+	0	–	–	–	–	–	–	–
Pearson χ^2		396.74	1.78	2.28	39.39	34.96	47.46	5.60
LRT χ^2		338.84	1.78	2.30	42.38	36.07	43.58	5.65
d.f		11	2	2	14	13	13	2

statistics, the Pearson statistic

$$\chi_P^2 = \sum \frac{\left(t_j - \hat{t}_j\right)^2}{\hat{t}_j}$$

and the likelihood-ratio statistic (LRT)

$$\chi_{LR}^2 = 2 \sum t_j \log \left(t_j / \hat{t}_j\right).$$

were calculated for each model. Both statistics were modified in the obvious way when categories were collapsed. The results are summarized in tables 1 and 2.

Our first remark is that the LSD performs badly with this data set, both in terms of fitting to the data as measured by the χ^2 statistics and in predicting the risk ratio and the total number of population uniques. This is not surprising as it agrees with the results of previous studies, e.g. Skinner et al. (1994), Chen and Keller-McNulty (1998) and Hoshino (2001). The fitted values of t_1 and t_2 show a poor fit to the observed values and the decay of the right hand tail appears to rapid. The resulting estimates of R_2 and T_1 are accordingly not satisfactory.

The PLN and the PiG models, save PiG (1), on the other hand both appear to adapt better to the frequency structure. The poor results of the PiG (1) is apparently a result of ignoring the large number of empty cells and assuming that $\theta = 0$. It appears as if most of the effort in fitting to the data is waisted on strectching out to t_0 on the expense of the other cell size frequencies. Even so, compared to the LSD even the PiG (1) model performs surprisingly well. When the zero-truncated methods are used better results are obtained both in fit to the data and in predicting R_2 and T_1. It is interesting to note that the best results with respect to predicting R_2 and T_1 are obtained when the estimation procedure is focused on the small cell sizes as in PiG (3) which is the PF12 estimation method. This seems to corroborate with the results of Chen and Keller-McNulty (1998). Furthermore, as mentioned in the preceding subsection, the censoring and trunction thresholds of the PLN (1) and (2) and the PiG (4) were opportunisticly chosen to produce estimates close to the true value of R_2. We found that higher thresholds for the PLN increasingly underestimated R_2 while for the PiG (4), R_2 was increasingly overestimated.

7 Remarks

Since the scope of this paper has been limited to a theoretical review with only a small-scale example, it is necessarily difficult to evaluate how the

PiG model fares in general and when compared to alternative approaches. Before any conclusions can be made a more extensive evaluation is called for including tests on real-life data and comparisons with other models. Such an evaluation is intended to appear in a separate report. The PiG model does however provide an analytically tractable alternative and calculations of the disclosure risk along the lines discussed are easily computed.

In the present paper we have only considered the two-parameter version of the more general three-parameter PiG, commonly known as the Sichel distribution. It is defined by

$$P_j = \frac{\eta^{-\gamma}}{j!}\left(\frac{\mu}{\eta}\right)^j \frac{K_{\gamma+j}\left(\mu\eta/\tau\right)}{K_\gamma\left(\mu/\tau\right)}.$$

where $-\infty < \gamma < \infty$. This distribution was introduced by Sichel (1971) and the distribution in (5) is obtained by setting $\gamma = -1/2$. A short review of the Sichel distribution is given e.g. in Johnson et al. (1992, pp. 455-457). This three-parameter distribution is very powerful and a number of known distribution functions such as the Poisson, negative binomial, geometric, Fisher's logarithmic series, are special or limiting forms of the Sichel. A problem with the Sichel distribution is however that the derivative $(\partial/\partial\gamma) K_\gamma(z)$ is not available in closed form and ML-estimation of γ requires special attention, see Stein et al. (1987).

We note also that the risk measures in (12) and (14) provide only an overall measure of disclosure risk pertaining to the sample as a whole. A per-record measure of disclosure risk is perhaps more useful as it would provide a means to identify sensitive (unique) records to which disclosure controlling measures can be applied. From an intruders point of view it would be optimal to utilize as much as possible of the information provided in the sample when formulating a model. Methods which attempt to capture the underlying probability structure inherent from the key variables defining X have been suggested, see e.g. Fienberg and Makov (1998) and Skinner and Holmes (1998). The latter considered a per-record measure based on their Poisson-lognormal model and we note that similar regression methods are available for PiG data based on a model of the form $\mu_{\mathbf{x}} = \exp\left(\mathbf{x}'\boldsymbol{\beta}\right)$ with τ fixed, see Dean et al. (1989) for details. Furthermore, as pointed out by an anonymous referee, the problem can also be addressed from a Bayesian viewpoint. In the Nordic European 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. In conclusion, the possibility of extending the present model in these directions is certainly worthy of future exploration.

8 Acknowledgment

The author wishes to thank Professor Daniel Thorburn, Dep. of Statistics, Stockholm University, for his many comments and suggestions in the preparation of this paper and Hans Block, Statistics Sweden, for providing the data set used in the example. The author also acknowledges the valuable comments of an anonymous referee.

A Derivation of Likelihood Equations

In the following the index s on the parameters, indicating sample level, is dropped for notational ease. The first derivatives of the log probabilities with respect to the parameters are (see Willmot, 1987, for details)

$$\frac{\partial \log p_j}{\partial \mu} = \frac{1}{\tau} + \frac{2j}{\mu} - \frac{\eta^2}{\mu\tau}\varphi_j \tag{27}$$

and

$$\frac{\partial \log p_j}{\partial \tau} = -\frac{\mu}{\tau}\left(\frac{\partial \log p_j}{\partial \mu}\right) + \frac{j}{\tau} - \frac{\varphi_j}{\tau} \tag{28}$$

where $\varphi_j = (j+1)\, p_{j+1} p_j^{-1}$ from which it in turn is easy to derive that

$$\frac{\partial p_j}{\partial \mu} = \frac{1}{\tau} p_j + \frac{2}{\mu} j p_j - \frac{\eta^2}{\mu\tau}(j+1)\, p_{j+1}$$

and

$$\frac{\partial p_j}{\partial \tau} = -\frac{\mu}{\tau}\left(\frac{\partial p_j}{\partial \mu}\right) + \frac{1}{\tau} j p_j - \frac{1}{\tau}(j+1)\, p_{j+1}.$$

Furthermore we need

$$\frac{\partial}{\partial \mu}\log(1-p_0) = -\frac{1-\eta}{\tau}\frac{p_0}{1-p_0}$$

and

$$\frac{\partial}{\partial \tau}\log(1-p_0) = -\frac{\mu}{\tau}\frac{1+\tau-\eta}{\tau\eta}\frac{p_0}{1-p_0}.$$

For the second derivatives we will need also the derivatives of φ_j and it is straightfoward using (27) and (28) to show that

$$\frac{\partial \varphi_j}{\partial \mu} = \frac{2}{\mu}\varphi_j - \frac{\eta^2}{\mu\tau}\varphi_j\left(\varphi_{j+1} - \varphi_j\right)$$

and

$$\frac{\partial \varphi_j}{\partial \tau} = -\frac{1}{\tau}\varphi_j + \frac{1+\tau}{\tau^2}\varphi_j\left(\varphi_{j+1} - \varphi_j\right).$$

In the following derivation of the likelihood equations we use the same line of arguments as Willmot (1987) and as an example we consider the zero-truncated conditional loglikelihood (CL) in (23); the other cases are analogous. The first derivatives of (23) with respect to μ and τ are

$$\begin{aligned} \frac{\partial l_{CL}}{\partial \mu} &= \sum_{j=1}^{\infty} t_j \frac{\partial \log p_j}{\partial \mu} - (C - t_0)\frac{\partial \log(1-p_0)}{\partial \mu} && \text{(29a)} \\ &= \frac{C-t_0}{\tau} + \frac{2n}{\mu} - \frac{\eta^2}{\mu\tau}\sum_{j=1}^{\infty} t_j\varphi_j + (C-t_0)\frac{(1-\eta)\,p_0}{\tau(1-p_0)} && \text{(29b)} \end{aligned}$$

and

$$\begin{aligned} \frac{\partial l_{CL}}{\partial \tau} &= \sum_{j=1}^{\infty} t_j \frac{\partial \log p_j}{\partial \tau} - (C - t_0)\frac{\partial \log(1-p_0)}{\partial \tau} && \text{(30a)} \\ &= -\frac{\mu}{\tau}\sum_{j=1}^{\infty} t_j \frac{\partial \log p_j}{\partial \mu} + \frac{n}{\tau} - \frac{1}{\tau}\sum_{j=1}^{\infty} t_j\varphi_j && \text{(30b)} \\ &\quad + (C-t_0)\frac{\mu(1+\tau-\eta)\,p_0}{\tau^2\eta(1-p_0)} \end{aligned}$$

respectively. It is clear that the partials of (23) with respect to μ and τ are identically zero when the likelihood is maximized, i.e. at the CL-estimates $\hat{\mu}$ and $\hat{\tau}$. Thus from (29a) we have that

$$\sum_{j=1}^{\infty} t_j \left.\frac{\partial \log p_j}{\partial \mu}\right|_{\mu=\hat{\mu},\tau=\hat{\tau}} = (C-t_0)\left.\frac{\partial \log(1-p_0)}{\partial \mu}\right|_{\mu=\hat{\mu},\tau=\hat{\tau}} \tag{31}$$

and it follows from setting (30b) equal to zero and using (31) that

$$\sum_{j=1}^{\infty} t_j\hat{\varphi}_j = n - \frac{\hat{\mu}(C-t_0)}{\hat{\eta}}\frac{\hat{p}_0}{1-\hat{p}_0} \tag{32}$$

where $\hat{\varphi}_j$, $\hat{\eta}$ and $\hat{p}_0$ are the CL-estimates of φ_j, η and p_0 respectively. Thus, setting (29) and (30) equal to zero and using (32) in (29b) yields after simplification the first likelihood equation h_1 in (24). The second equation h_2

is simply (32) with μ replaced by $n\left(1-p_0\right)\left(C-t_0\right)^{-1}$ from the first equation. It is straightfoward using the results above to show that the required derivatives of h_1 and h_2 are

$$\begin{aligned}
\frac{\partial h_1}{\partial \mu} &= \frac{1}{1-p_0}+\frac{\mu\left(1-\eta\right)}{\tau}\frac{p_0}{\left(1-p_0\right)^2} \\
\frac{\partial h_1}{\partial \tau} &= \frac{\mu^2\left(1+\tau-\eta\right)}{\tau^2\eta}\frac{p_0}{\left(1-p_0\right)^2} \\
\frac{\partial h_2}{\partial \mu} &= \frac{2}{\mu}\sum_{j=1}^{\infty}t_j\varphi_j-\frac{\eta^2}{\mu\tau}\sum_{j=1}^{\infty}t_j\varphi_j\left(\varphi_{j+1}-\varphi_j\right)-\frac{n\left(1-\eta\right)p_0}{\tau\eta} \\
\frac{\partial h_2}{\partial \tau} &= -\frac{1}{\tau}\sum_{j=1}^{\infty}t_j\varphi_j+\frac{1+\tau}{\tau^2}\sum_{j=1}^{\infty}t_j\varphi_j\left(\varphi_{j+1}-\varphi_j\right)+\frac{np_0}{\eta}\left(\frac{\mu\left(1+\tau-\eta\right)}{\tau^2\eta}-\frac{1}{\eta^2}\right).
\end{aligned}$$

References

[1] Abramovitz, M. and Stegun, I.A. (1970) *Handbook of Mathematical Functions.* Monograph. New York: Dover Publications.

[2] Bethlehem, J.G., Keller, W.J. and Pannekoek, J. (1990) Disclosure Control of Microdata. *Journal of the American Statistical Association*, **85**, pp. 38-45.

[3] Blien, U., Müller, W. and Wirth, H. (1993) Needles in Haystacks Are Hard to Find - Testing Disclosure Risks of Anonymous Individual Data. In *Proceedings of the International Seminar on Statistical Confidentiality, Dublin, 8-10 September 1992*, pp. 391-406. Luxembourg: Office for Official Publications of the European Communities.

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

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

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

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

[8] 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.

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

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

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

[12] 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.

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

[15] 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.

[16] 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*, pp. 175-187. Swedish Council for Social Science Research.

[17] Frank, O. (1988) Designing Classifiers for Partial Information Release, in H.H. Bock (ed.) *Classification and Related Methods of Data Analysis: Proceedings of the First Conference of the IFCS, Tech. Univ. of Aachen,* pp. 687-690. New York: North-Holland.

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

[19] Holla, M.S. (1966) On a Poisson-inverse Gaussian Distribution, *Metrika*, **11**, pp. 115-121.

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

[21] 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.

[22] Johnson, N.L., Kotz, S. and Balakrishnan, N. (1994) *Continuous Univariate Distributions*, Vol. 1, 2nd ed.. Monograph. New York: John Wiley & Sons.

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

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

[25] Matlab (2001) Matlab Ver. 6.1, Release 12.1, MathWorks Inc..

[26] 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.

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

[28] 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.

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

[30] Sichel, H.S. (1971) On a Family of Discrete Distributions Particularly Suited to Represent Long-tailed Frequency Data. In N.F. Laubscher (ed.) *Proceedings of the Third Symposium on Mathematical Statistics*, pp. 51-97. Pretoria: C.S.I.R.

[31] Sichel, H.S. (1973) The Density and Size Distribution of Diamonds, *Bulletin of the International Statistical Institute*, **45**(2), pp. 420-427.

[32] 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.

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

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

[35] Sichel, H.S. (1982b) Asymptotic Efficiencies of Three Methods of Estimation for the Inverse Gaussian-Poisson Distribution, *Biometrika*, **69**, pp. 467-472.

[36] 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.

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

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

[39] 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.

[40] Stein, G., Zucchini, W. and Juritz, J.M. (1987) Parameter Estimation for the Sichel Distribution and its Multivariate Extension, *Journal of the American Statistical Association*, **82**, pp. 938-944.

[41] 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.

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

[43] 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.

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