On a Conditional Inverse Gaussian-Poisson Distribution

Nobuaki Hoshino[1]

December 2001

CMU-CALD-01-102

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

The present article treats a distribution of random partitioning of the positive integer. Although such a distribution is important concerning applications in many fields such as statistical ecology, linguistics and statistical disclosure control, not very many models are known owing to the difficulty caused by inevitable combinatorics. The present article shows that conditioning the total frequency of an inverse Gaussian-Poisson population model leads to a new manipulatable distribution of random clustering. We also give formulae that are necessary for the application of this distribution.

[1]Visiting faculty from Faculty of Economics, Kanazawa University. *Address for correspondence* : Nobuaki Hoshino, Faculty of Economics, Kanazawa University, Kakuma-machi, Kanazawa-shi, Ishikawa, Japan. E-mail: hoshino@kenroku.kanazawa-u.ac.jp

This manuscript is written during the author's visit to Carnegie Mellon University, which is financed by the Japanese ministry of education, culture, sports, science and technology. The author would like to express sincere thanks for their support.

Keywords: Random clustering, Species abundance, Superpopulation, disclosure risk

1 Introduction

Scientists observe various kinds of populations. In many instances a population consists of diverse groups, and its property is hard to formulate. To comprehend the complex nature of a population, it is often useful to focus upon its heterogeneity. This is a classical theme in statistics, and we can for example date back to Neyman (1939). We shall later see more examples, in which the measurement of heterogeneity possesses great importance. Statistically we have utilized population models that are suited to such measurement, and the present article proposes another population model.

In Section 1.1, we review the background of statistical population models. Section 1.2 derives the model that we propose. Some theoretical results on this model are shown in Section 2, and we discuss parameter estimation in Section 3. Finally Section 4 provides application results and concluding remarks.

1.1 A population model

Let us consider the following population model: the total number of cells equals J, and $F_j, j = 1, \ldots, J$, are the size of the j-th cell; the population size is denoted by $N = \sum_{j=1}^{J} F_j$. Let S_i denote the number of cells of size i. More specifically,

$$S_i = \sum_{j=1}^{J} I(F_j = i), \quad i = 0, 1, \ldots$$

where $I(\cdot)$ is the indicator function:

$$I(F_j = i) = \begin{cases} 1, & F_j = i, \\ 0, & F_j \neq i. \end{cases}$$

In literatures, $(S_0, S_1, \ldots)$ are called size indices (Sibuya (1993)), frequencies of frequencies (Good (1965)) or equivalence class (Greenberg and Zayatz (1992)).

Obviously

$$\sum_{i=0}^{\infty} S_i = J, \qquad \sum_{i=1}^{\infty} i \cdot S_i = N.$$

Note that J is the total number of cells including the number of the empty cells S_0. In the following we denote the number of non-empty cells by

$$U = J - S_0 = \sum_{i=1}^{\infty} S_i.$$

Many authors have regarded F_j's as random variables. Under such an assumption, we can summarize the information of a population with only at most a few parameters. For example, Fisher et al. (1943) developed the logarithmic series distribution to summarize a population of Malayan butterflies, which commenced the vast studies of statistical ecology or stochastic abundance models. In this situation, a population is composed of J species, and the number of j-th species corresponds to F_j. See Engen (1978) for the context. In addition, there are myriads of examples in linguistics. A writer is deemed to have a vocabulary of J words, and each F_j

corresponds to the frequency of the usage of j-th word in the writer's text. Williams (1956) provides an early review of this context.

In these applications, statistical interest lies in the χ^2 test that determines whether one can regard data as being subject to an assumed distribution, since the summary of that kind is meaningless under an ill-assumed distribution. Thus most of researchers have investigated skew distributions that fit empirically well to data in the sense of χ^2. As a result, their purpose is just fitting to sample data, and there is little interest in the structure of the corresponding population.

However, in some cases the objective of an analysis is to estimate the population structure about size indices. Obviously some ecologists are interested in not samples but a whole population. Let us mention other examples. When a statistical agency disseminates microdata, it is very important to measure the risk of privacy invasion. An individual that is unique in a population is considered to be unsafe to publish. Thus S_1 is a typical index of the risk, and its estimation is necessary unless data are of a census. See Willenborg and de Waal (1996) for the context of statistical disclosure control. Also we can find a similar problem in database merging. When databases have common individuals, it is necessary to identify how many individuals are in common. Since a database is often composed of numerous records, it is valuable to estimate such overlaps based on samples.

As regards the estimation of size indices, it is useful to assume a random population model, even though the objective is not to summarize information. Let us assume simple random sampling without replacement; if there is no assumption about a finite population, the unique unbiased estimator of S_i, if any, is useless because of its large variance. See Section 2.3 of Engen (1978). The author would rather adopt the superpopulation model approach or the empirical Bayes method. Namely the estimator of a size index is its expectation under F_j given the estimates of the parameters of F_j. For example, Bethlehem et al. (1990) regarded F_j's as gamma-Poisson mixture, which has two parameters; the parameters were estimated from data, whence they calculated $\mathrm{E}(S_1)$ as an index of the risk.

The present article assumes the superpopulation approach to estimate population size indices. As we have mentioned, much attention has not been paid to the relationship between samples and the population. Therefore we should clarify that our approach needs to treat explicitly such a relationship. In the following the sample size is denoted by n. Sample size indices are similarly defined and denoted by $(s_0, s_1, \ldots)$. The number of non-empty cells is $u = \sum_{i=1} s_i$. Assuming a population model, we can derive the sampling distribution in terms of $(s_0, \ldots)$. Then we construct estimators of the parameters of the population model. The estimator of S_i is its expectation given the estimates of parameters, and $\mathrm{E}(S_i)$ depends on the total frequencies $N = \sum_j F_j$.

In many cases, including Bethlehem et al. (1990), F_j's are regarded as independently identically distributed Poisson mixture. Then the population size N is a random variable, and this may be a problem because concerning the estimation of size indices the population size is given and fixed. To ease this conflict, one may assume $\mathrm{E}(N) = N_0$ as in Bethlehem et al. (1990), where the number of parameters was reduced to one. Under the restriction, we can immediately obtain the sampling distribution with Bernoulli sampling (Särndal et al. (1992)), in which each individual is independently sampled with success probability n_0/N_0. Namely, the sampling distribution is the result of substituting N_0 in the distribution of a population by the observed number of samples n_0. Such treatment is thus expedient.

However, simple random sampling without replacement requires the population size N to be fixed at N_0. Hence population models where $N = N_0$ are more realistic in cases such as statistical disclosure control, where simple random sampling without replacement is employed. Although Bernoulli sampling may be valid as an approximation of simple random sampling without replacement, we ought to investigate models in which $N = N_0$ also.

The difficulty of a model where its population size is fixed is that it involves combinatorics. In fact it is equivalent to random partitioning of the positive integers, which is itself an interesting subject of probability; see Hoshino (2001) for an application of a formula of this field. Although we can utilize existing results, only a few models are treatable to the author's knowledge. Hence there is great need to develop useful models that satisfy the size restriction, in order to handle various populations.

Among known models, the Dirichlet-multinomial model is obtained by conditioning the gamma-Poisson model as $N = N_0$ (Takemura, 1999). Similarly, if we can easily derive the distribution of N under independently identically distributed F_j's, the construction of a size-restricted model may be straightforward. Holla (1966) introduced inverse Gaussian-Poisson mixture, which is closed under convolution; the present article investigates the conditional population model of inverse Gaussian-Poisson mixture.

The inverse Gaussian-Poisson mixture is a special case of the generalized inverse Gaussian-Poisson mixture proposed by Sichel (1971), which is, however, less treatable than the inverse Gaussian-Poisson mixture. See Jørgensen (1982) for the generalized inverse Gaussian distribution. Concerning the (generalized) inverse Gaussian-Poisson mixture, there are a certain number of applications in statistical ecology and linguistics. Here we only mention Sichel (1997) as an example, though his population model is different from ours. Seshadri (1999) provides an excellent review on the inverse Gaussian distribution; its Section 7.1 is devoted to the inverse Gaussian-Poisson mixture. Since the inverse Gaussian-Poisson mixture has been used to describe populations, our approach seems to be promising about applications.

1.2 The derivation of a conditional inverse Gaussian-Poisson distribution

The inverse Gaussian distribution is defined as

$$\mathrm{P}(\lambda; \alpha, \theta) = \frac{(2\sqrt{1-\theta}/(\alpha\theta))^{\frac{-1}{2}}}{2K_{-1/2}(\alpha\sqrt{1-\theta})} \lambda^{-\frac{3}{2}} \exp(-(\frac{1}{\theta} - 1)\lambda - \frac{\alpha^2\theta}{4\lambda}) \tag{1}$$

for $0 < \theta < 1, \alpha > 0$, where

$$K_{-1/2}(\xi) = \sqrt{\frac{\pi}{2}} \xi^{-1/2} \exp(-\xi) \tag{2}$$

is the modified Bessel function of the third kind of order $-1/2$.

Suppose that a random variable Y is distributed as Poisson with mean λ, and let λ distribute with its density (1). Then the distribution of Y is widely known to be

$$\mathrm{P}_{IGP}(Y = y; \alpha, \theta) = \sqrt{\frac{2\alpha}{\pi}} \exp(\alpha\sqrt{1-\theta}) \frac{(\alpha\theta/2)^y}{y!} K_{y-1/2}(\alpha), \quad y = 0, 1, 2, \ldots. \tag{3}$$

See Chapter 7.1 of Seshadri (1999) for more detail. In the present article, we refer to (3) as the inverse Gaussian-Poisson distribution and denote it by $IGP(\alpha, \theta)$.

In fact we can obtain

$$K_{y-1/2}(\alpha) = \sqrt{\frac{\pi}{2\alpha}}\exp(-\alpha)(\sum_{i=0}^{y-1}\frac{(y-1+i)!}{(y-1-i)!i!}(2\alpha)^{-i}), \quad y = 1, 2, \ldots,$$

from (2) and $K_{-1/2}(\xi) = K_{1/2}(\xi)$, by means of

$$K_{\gamma+1}(\alpha) = \frac{2\gamma}{\alpha}K_{\gamma}(\alpha) + K_{\gamma-1}(\alpha).$$

In addition, Ismail (1977) showed that

$$K_{\gamma}(\alpha) \approx 2^{\gamma}\gamma^{\gamma}\exp(-\gamma)\alpha^{-\gamma}\sqrt{\frac{\pi}{2\gamma}} \tag{4}$$

when γ is large. Equation (4) is useful because the modified Bessel function of the third kind may overflow as $\gamma \to \infty$. Consequently, computation on the inverse Gaussian-Poisson distribution is not very hard. Consult Watson (1944) for the results of Bessel functions.

Henceforth we consider the population model that is discussed in Section 1.1 under the assumption that $F_j, j = 1, \ldots, J$ are independently identically distributed as $IGP(\alpha, \theta)$. Namely, we suppose that

$$\mathrm{P}(F_1, \ldots, F_J) = \prod_{j=1}^{J}\sqrt{\frac{2\alpha}{\pi}}\exp(\alpha\sqrt{1-\theta})\frac{(\alpha\theta/2)^{F_j}}{F_j!}K_{F_j-1/2}(\alpha),$$

or

$$\mathrm{P}(S_0, \ldots) = J!\prod_{i=0}^{\infty}\{\sqrt{\frac{2\alpha}{\pi}}\exp(\alpha\sqrt{1-\theta})\frac{(\alpha\theta/2)^{i}}{i!}K_{i-1/2}(\alpha)\}^{S_i}\frac{1}{S_i!}. \tag{5}$$

One merit of the IGP model (5) is that we can evaluate the exact distribution of the population size N. The probability generating function of (3) is shown as

$$G(z) = \exp(\alpha(\sqrt{1-\theta} - \sqrt{1-z\theta})) \tag{6}$$

by Sankaran (1968). Hence we can see that the sum of J random variables that are independently identically distributed as $IGP(\alpha, \theta)$ is distributed as $IGP(J\alpha, \theta)$. That is,

$$\mathrm{P}(N) = \sqrt{\frac{2J\alpha}{\pi}}\exp(J\alpha\sqrt{1-\theta})\frac{(J\alpha\theta/2)^{N}}{N!}K_{N-1/2}(J\alpha). \tag{7}$$

We are interested in the conditional population model given its population size N; the model (5) divided by (7) becomes

$$\begin{aligned}\mathrm{P}(S_0, \ldots, S_N|N) &= \frac{J!\prod_{i=0}^{N}\{\sqrt{\frac{2\alpha}{\pi}}\exp(\alpha\sqrt{1-\theta})\frac{(\alpha\theta/2)^{i}}{i!}K_{i-1/2}(\alpha)\}^{S_i}\frac{1}{S_i!}}{\sqrt{\frac{2J\alpha}{\pi}}\exp(J\alpha\sqrt{1-\theta})\frac{(J\alpha\theta/2)^{N}}{N!}K_{N-1/2}(J\alpha)} \\ &= (\frac{2\alpha}{\pi})^{\frac{J-1}{2}}\frac{J!N!}{J^{N+1/2}K_{N-1/2}(J\alpha)}\prod_{i=0}^{N}\{\frac{K_{i-1/2}(\alpha)}{i!}\}^{s_i}\frac{1}{s_i!}. \end{aligned} \tag{8}$$

The right hand side of (8) seems to be a new distribution with one parameter; it is worthy of note that (8) is derived from the distribution with two parameters. We refer to (8) as the Conditional Inverse Gaussian-Poisson distribution ($CIGP(\alpha)$).

2 On the property of $CIGP(\alpha)$

In this section we will clarify a few properties of $CIGP(\alpha)$, which are important in applications. We rewrite the right hand side of (8) as

$$\mathrm{P}_J(S_0,\ldots,S_N) = (\frac{2\alpha}{\pi})^{\frac{J-1}{2}} \frac{J!N!}{J^{N+1/2}K_{N-1/2}(J\alpha)} \prod_{i=0}^{N} \{\frac{K_{i-1/2}(\alpha)}{i!}\}^{S_i} \frac{1}{S_i!}, \tag{9}$$

where we explicitly denote the dependence of $CIGP(\alpha)$ on J. Note that $\alpha > 0$.

First we see relationships among distributions connected with the CIGP model. When we assume $\theta = 1$, the density of inverse Gaussian (1) equals the density of the reciprocal gamma distribution. Since the derivation of $CIGP(\alpha)$ does not depend on the value of θ, the conditional model of reciprocal gamma-Poisson mixture given N is $CIGP(\alpha)$. Takemura (1999) clarified that the conditional model of gamma-Poisson mixture (=negative binomial) given N equals Dirichlet-multinomial mixture, which is a multivariate generalization of beta-binomial mixture. Therefore, in a sense, the CIGP model corresponds to Dirichlet-multinomial mixture. See Hoshino and Takemura (1998) for more detailed discussion on distributions relating to gamma-Poisson mixture.

Then we show the expectations of size indices.

Theorem 1 *Suppose that size indices are distributed as* (9). *Then the factorial moments are*

$$\mathrm{E}\,(\prod_{j=1}^{N} S_j^{(r_j)}) = (\frac{2\alpha}{\pi})^{\frac{r}{2}} \frac{N!J!K_{N-R-1/2}((J-r)\alpha)(J-r)^{N-R+1/2}}{(N-R)!(J-r)!J^{N+1/2}K_{N-1/2}(J\alpha)} \prod_{j=1}^{N} (\frac{K_{j-1/2}(\alpha)}{j!})^{r_j},$$

where $r = \sum_{j=1}^{N} r_j$, $R = \sum_{j=1}^{N} j r_j$, *and* $S_j^{(r_j)} = S_j(S_j-1)\cdots(S_j-r_j+1)$.

Proof For simplicity, here we evaluate $E(S_j)$. Let us write

$$\mathcal{S}(N) = \{\boldsymbol{S} = (S_1,\ldots,S_N) | \sum_{i\geq 1} iS_i = N\}.$$

Note that for $j = 1, 2, \ldots$, if $S_j \geq 1$ then

$$\begin{aligned} S_j\,\mathrm{P}_J(S_0,\ldots,S_N|N) &= \mathrm{P}_{J-1}(S_0,\ldots,S_{j-1},S_j-1,S_{j+1},\ldots,S_N|N-j) \\ &\quad \times\sqrt{\frac{2\alpha}{\pi}}(\frac{K_{j-1/2}(\alpha)}{j!})\frac{N!K_{N-j-1/2}((J-1)\alpha)(J-1)^{N-j+1/2}}{(N-j)!J^{N-1/2}K_{N-1/2}(J\alpha)}, \end{aligned}$$

else $S_j = 0$ and $S_j\,\mathrm{P}_J(S_0,\ldots,S_N|N) = 0$. Thus

$$\begin{aligned} \mathrm{E}_J(S_j|N) &= \sum_{\boldsymbol{S}\in\mathcal{S}(N)} S_j\,\mathrm{P}_J(S_0,\ldots,S_N|N) \\ &= \sqrt{\frac{2\alpha}{\pi}}(\frac{K_{j-1/2}(\alpha)}{j!})\frac{N!K_{N-j-1/2}((J-1)\alpha)(J-1)^{N-j+1/2}}{(N-j)!J^{N-1/2}K_{N-1/2}(J\alpha)} \\ &\quad \times \sum_{\boldsymbol{S}\in\mathcal{S}(N)} \mathrm{P}_{J-1}(S_0,\ldots,S_{j-1},S_j-1,S_{j+1},\ldots,S_N|N-j)\; I(S_j\geq 1). \end{aligned} \tag{10}$$

Because

$$\sum_{\boldsymbol{S}\in\mathcal{S}(N)} \mathrm{P}_{J-1}(S_0,\ldots,S_{j-1},S_j-1,S_{j+1},\ldots,S_N|N-j)\, I(S_j\geq 1)$$
$$= \sum_{\boldsymbol{S}\in\mathcal{S}(N-j)} \mathrm{P}_{J-1}(S_0,\ldots,S_{N-j}|N-j) = 1,$$

we obtain

$$\mathrm{E}_J(S_j|N) = \sqrt{\frac{2\alpha}{\pi}}\left(\frac{K_{j-1/2}(\alpha)}{j!}\right)\frac{N!K_{N-j-1/2}((J-1)\alpha)(J-1)^{N-j+1/2}}{(N-j)!J^{N-1/2}K_{N-1/2}(J\alpha)} \tag{11}$$

from (10). Similarly, we can show the theorem. Q.E.D.

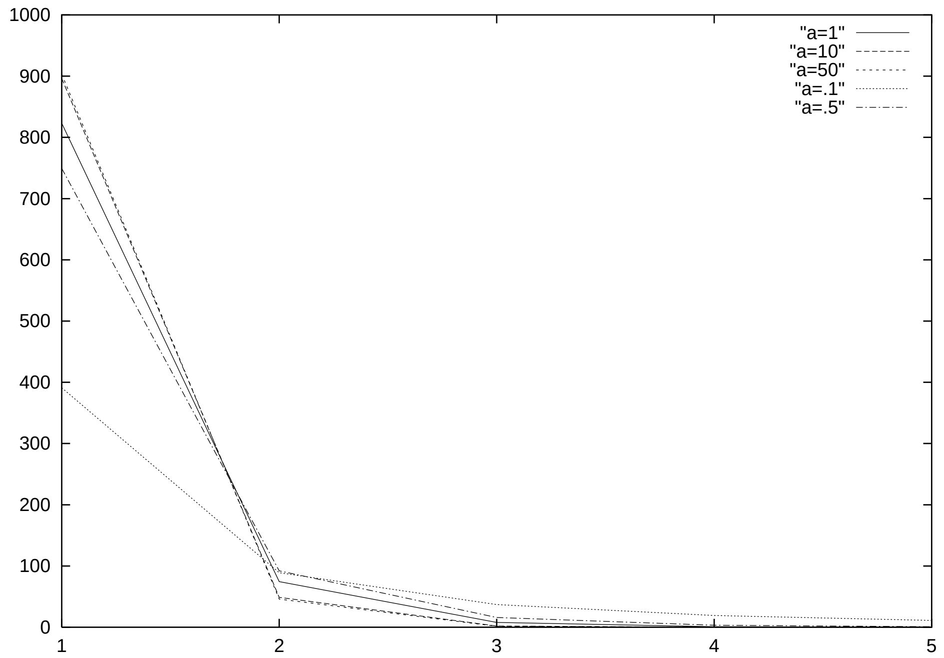

Figure 1: *The expectations of size indices with various parameter values of α*

Figure 1 plots the expectations of size indices with parameter values $\alpha = 0.1, 0.5, 1, 10, 50$ under $N = 1000$ and $J = 10000$. The vertical axis shows $\mathrm{E}(S_j)$, and the horizontal axis corresponds to $j = 1, 2, \ldots, 5$. We can observe that the expectation of a size index is decreasing with respect to the size; this is the pattern that we frequently find in applications. The difference between the size indices of $\alpha = 10$ and $\alpha = 50$ appears rather small compared to the difference between the values of $\alpha = 0.1$ and $\alpha = 0.5$.

Now we discuss the sampling distribution of $CIGP(\alpha)$. Because the CIGP model does not depend on the label of each individual, Lemma 1 of Takemura (1999) assures that we can immediately derive the sampling distribution as a result of substituting N of the population distribution by n. In other words, the distribution of n samples directly drawn from the infinite population ($CIGP(\alpha)$) is the same as that of n samples from the finite population of size N that is from the infinite population.

Theorem 2 *Suppose that the distribution of population size indices is (9) and n samples are drawn with simple random sampling without replacement. Then the sample size indices are distributed according to*

$$P_J(s_0,\ldots,s_n) = (\frac{2\alpha}{\pi})^{\frac{J-1}{2}} \frac{J!n!}{J^{n+1/2}K_{n-1/2}(J\alpha)} \prod_{i=0}^{n} \{\frac{K_{i-1/2}(\alpha)}{i!}\}^{s_i} \frac{1}{s_i!}. \tag{12}$$

3 Parameter estimation

This section treats the estimation of α from samples that are subject to $CIGP(\alpha)$. We can provide the maximum likelihood estimator and two approximate estimators.

3.1 Maximum Likelihood (ML) estimation

We will denote the log likelihood of (12) by

$$L = \frac{J-1}{2}\log(2\alpha) - \log K_{n-1/2}(J\alpha) + \sum_{i=0}^{n} s_i \log K_{i-1/2}(\alpha) + Const.$$

In the following we will use this notation:

$$R_\gamma(\alpha) = \frac{K_{\gamma+1}(\alpha)}{K_\gamma(\alpha)},$$

and it is widely known that

$$\frac{\partial \log K_\gamma(\alpha)}{\partial \alpha} = -R_\gamma(\alpha) + \frac{\gamma}{\alpha}. \tag{13}$$

See Seshadri (1999, p.125) for instance.

Now we construct the ML estimator: the solution of $dL/d\alpha = 0$. Equation (13) leads to the following expression of the derivative of L:

$$\begin{aligned} \frac{dL}{d\alpha} &= \frac{J-1}{2\alpha} - \{-R_{n-1/2}(J\alpha) + \frac{n-1/2}{J\alpha}\}J + \sum_{i=0} s_i\{-R_{i-1/2}(\alpha) + \frac{i-1/2}{\alpha}\} \\ &= JR_{n-1/2}(J\alpha) - \sum_{i=0}^{n} s_i R_{i-1/2}(\alpha), \end{aligned}$$

by

$$\sum_{i=0}^{n} s_i = J \quad \text{and} \quad \sum_{i=0}^{n} i s_i = n.$$

The ML estimate obviously requires numerical evaluation; we can adopt the Newton-Raphson method using the second derivative:

$$\frac{d^2L}{d\alpha^2} = J^2\{R_{n-1/2}^2(J\alpha) + \frac{2n}{J\alpha}R_{n-1/2}(J\alpha)\} - \sum_{i=0} s_i\{R_{i-1/2}^2(\alpha) + \frac{2i}{\alpha}R_{i-1/2}(\alpha)\} - J^2 + J.$$

Note that

$$\frac{\partial R_{\gamma-1/2}(\alpha)}{\partial \alpha} = R_{\gamma-1/2}^2(\alpha) - \frac{2\gamma}{\alpha}R_{\gamma-1/2}(\alpha) - 1.$$

The estimators that are discussed in Section 3.2 can be used for the starting value of α in such an iteration procedure.

3.2 Approximate estimation

The expectation of a sample size index is derived from substituting N of (11) by n. Therefore we may be able to apply the method of moments. However, what is inconvenient is the evaluation of polynomials of order n (i.e. $K_{n-1/2}(\cdot)$). Here we utilize estimators of the IGP distribution to construct an approximate moment estimator of the CIGP parameter, because it is easy to calculate. We also introduce another approximate estimator, which is a function of s_0.

If Y is a random variable that is subject to $IGP(\alpha, \theta)$, equation (6) implies that

$$\mathrm{E}(Y) = \frac{\alpha\theta}{2\sqrt{1-\theta}},$$

and

$$\mathrm{V}(Y) = \frac{\alpha\theta(2-\theta)}{4(1-\theta)^{\frac{3}{2}}}.$$

We substitute $\mathrm{E}(Y)$ by sample average n/J, and $\mathrm{V}(Y)$ by sample variance:

$$v = \frac{\sum_{i=0}^{n}(i - n/J)^2 s_i}{J}.$$

The solution of these simultaneous equations is given by

$$\theta = \frac{2n - 2Jv}{n - 2Jv}, \quad \text{and} \quad \alpha = \frac{2n\sqrt{1-\theta}}{J\theta}.$$

We propose using above equation about α as our approximate estimator:

$$\tilde{\alpha} = \frac{n\sqrt{n(2Jv - n)}}{J(Jv - n)}. \tag{14}$$

Although (14) is simple enough, its efficiency may not be enough. On the IGP distribution, Sichel (1982) calculated asymptotic efficiencies for the joint estimation of α and θ for the method of moments. According to his result, the method of moments is inefficient at α being as great as 10 when $\theta = 0.97$, which is claimed to be a typical case of parameter values. Sichel (1973) proposed another estimator of the IGP parameter α, whose efficiency was high for small α in the Sichel (1982)'s experiment. It leads to

$$\bar{\alpha} = -\frac{1}{2}(\log s_0 - \log J)(1 + \frac{n/J}{n/J + \log s_0 - \log J}) \tag{15}$$

in our setting. See Section 4 for an empirical comparison of these estimators.

4 Application results and conclusions

In this section we examine the applicability of the CIGP model to real data. We fit the CIGP model to plankton data (Table 1), lice data (Table 2) and Japanese labor force survey data (Table 3). The present article then concludes with some remarks.

Barnes and Marshall (1951) provided plankton data series; Reid (1981) fitted Log-Normal-Poisson mixture (LNP, 2 parameters), Gamma-Poisson mixture (GP, 2 parameters) and Neyman type A (NY, 2 parameters) distribution to a data set of $n = 232$ and $J = 120$ from the series. In these models, the total frequency is not fixed. Namely, n is the sum of independently identically distributed random variables. Now we apply $CIGP(\alpha)$ to the same data set; the results are shown in Table 1. Concerning the CIGP model, the ML estimate appears to be $\hat{\alpha} = 10.35$ in this case, and the fitted values of size indices are the expectations under $\hat{\alpha}$. The approximate moment estimate by (14) is $\tilde{\alpha} = 4.49$, and another estimate by (15) is $\bar{\alpha} = 6.50$; these estimates are not very close to the ML estimate. We observe that the fits of the models are satisfactory in terms of the χ^2 criterion. If n is regarded as known, it is reasonable to include the information of the sample size in a model, whereby the degree of freedom increases without great loss of fit as the case of the CIGP model.

Next we compare the CIGP model with the IGP model (3). Stein et al. (1987) fitted the IGP distribution to lice data (William, 1964) wherein $n = 7442, J = 1083$. We apply the CIGP model to the same data; see Table 2 for the results. These fits are bad, yet similar. We also observe that ML estimates $\hat{\alpha}_{IGP} = 0.645$ and $\hat{\alpha}_{CIGP} = 0.644$ are similar. According to Sichel (1982), α_{IGP} describes the shape of the distribution, whereas θ_{IGP} controls the upper tail. This seems to explain the similarity between $\hat{\alpha}_{IGP}$ and $\hat{\alpha}_{CIGP}$. The approximate moment estimate is $\tilde{\alpha} = 1.069$, and another estimate is $\bar{\alpha} = 0.579$ here.

Sichel (1982) evaluated the asymptotic covariance matrix of $\hat{\alpha}_{IGP}$ and $\hat{\theta}_{IGP}$. He remarked that the correlation between $\hat{\alpha}_{IGP}$ and $\hat{\theta}_{IGP}$ is negative, and is generally substantial in the useful range of values. Hence Stein et al. (1987) proposed a reparameterization to avoid numerical instability. However, as far as the CIGP model is concerned, n seems to determine the tail, which is governed by θ in the IGP model. We can thus regard the CIGP model as one way to overcome aforementioned numerical instability without the arbitrariness of a reparameterization.

We then demonstrate the applicability of the CIGP model to the estimation of population size indices in the field of statistical disclosure control. This is interesting because there seems to exist no application of the IGP distribution in this field. Sai and Takemura (2000) calculated the size indices of Japanese labor force survey data that were collected in December 1997. We apply the CIGP model to their anonymized data of Akita prefecture; our interest lies in the number of population uniques (S_1) with respect to the degree of the anonymization. Each record contains the information of variables such as sex or age; these variables are classified in some categories, with the result that J is the product of the number of categories in the variables. In this case, $J = 5.644 \times 10^{12}$ and $n = 908$. Observe Table 3 for the result of fitting. The ML estimate $\hat{\alpha}$ is 9.047×10^{-10}; $\tilde{\alpha} = 7.423 \times 10^{-10}$ and $\bar{\alpha} = 9.061 \times 10^{-10}$. The numbers of nonzero-frequency groups (u) are the same between the observed set and the fitted set.

Under the superpopulation approach, the estimator of S_1 is $\mathrm{E}(S_1|N)$ of the CIGP distribution, where N equals 1.028 million. However, the author could not compute the value within a reasonable time. Here we only give an approximate value $E(S_1) \approx 2553$; this can be obtained with the following proposition, which is an immediate consequence of (4) and (11).

Proposition 1 *Suppose that size indices are distributed according to (9). Then, as $N \to \infty$,*

$$\mathrm{E}(S_1|N) \approx \exp(1-\alpha)\frac{N\alpha(J-1)}{2}\frac{(N-3/2)^{N-2}}{(N-1/2)^{N-1}}.$$

Now we conclude the discussion with some remarks. The CIGP model can surely be used in the estimation of population size indices, and it has merits particularly in being free from the numerical instability that is reported on the ML estimation of the IGP parameters. The

i	s_i	LNP	GP	NY	CIGP
0	23	23.3	21.0	21.4	20.6
1	28	34.0	33.1	32.7	33.3
2	34	27.9	29.2	28.9	29.5
3	17	17.3	18.9	18.9	19.0
4	8	9.7	10.1	10.2	10.0
5	7	4.6	4.7	4.7	4.6
6	3	2.1	1.9	2.0	1.9
7+	0	1.1	1.1	1.1	1.1
$\chi^2(d.f.)$		5.44(5)	5.26(5)	5.09(5)	5.41(6)

Table 1: *Frequency distribution of Oithona similis nauplii (Barnes and Marshall, 1951)*

estimate by (15) tends to be closer to the ML estimate in our experiments, which suggests that $\bar{\alpha}$ may be better than $\tilde{\alpha}$ on real data. According to Takemura (1999), we can derive the Ewens distribution (See Chap. 41 of Johnson et al. (1997)) from Dirichlet-multinomial mixture by a limiting argument. The same kind of limiting distribution of the CIGP model, where $J\alpha$ is fixed and $\alpha \to 0$, will be discussed in the author's subsequent paper. The CIGP model needs the information of s_0, but many data in applications have no information of s_0. In such a case, we may be able to use that limiting distribution.

References

[1] Barnes, H. and Marshall, S.M. (1951). On the variability of replicate plankton samples and some applications of contagious series to the statistical distribution of catches over restricted periods. *Journal of the Marine Biological Association of U.K.*, **30**, 233–263.

[2] Bethlehem, J.G., Keller, W.J. and Pannekoek, J. (1990). Disclosure control of microdata. *Journal of the American Statistical Association*, **85**, 38–45.

[3] Fisher, R.A., Corbet, A.S. and Williams, C.B. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *Journal of Animal Ecology*, **12**, 42–58.

[4] Good, I.J. (1965). *The Estimation of Probabilities: An Essay on Modern Bayesian Methods.* MIT Press, Cambridge, Massachusetts.

[5] Greenberg, B.V. and Zayatz, L.V. (1992). Strategies for measuring risk in public use microdata file. *Statistica Neerlandica*, **46**, 33–48.

[6] Holla, M.S. (1966). On a Poisson-inverse Gaussian distribution. *Metrika*, **11**, 115–121.

[7] Hoshino, N. and Takemura, A. (1998). Relationship between logarithmic series model and other superpopulation models useful for microdata disclosure risk assessment. *Journal of the Japan Statistical Society*, **28**, 2, 125–134.

Lice per head	Number of heads	IGP	CIGP
0	622	585.50	585.70
1	106	188.49	188.18
2	50	77.36	77.18
3	29	41.85	41.75
4	33	26.77	26.71
5	20	18.91	18.87
6	14	14.25	14.22
7	12	11.22	11.20
8	18	9.12	9.10
9	11	7.60	7.59
10	11	6.45	6.45
11-12	13	10.44	10.43
13-14	14	8.13	8.13
15-16	9	6.56	6.56
17-18	11	5.43	5.44
19-21	17	6.63	6.64
22-24	12	5.33	5.34
25-28	15	5.70	5.71
29-33	11	5.57	5.59
34-40	15	5.91	5.93
41-48	13	5.03	5.05
49-60	8	5.45	5.49
61-76	4	5.00	5.05
77-102	4	5.23	5.29
103+	11	15.15	15.30
$\hat{\alpha}$ by MLE		0.645	0.644
$\hat{\theta}$ by MLE		0.998	

Table 2: *Frequency distribution of Lice (Williams, 1964)*

i	1	2	3	4	5	6	7+	u
s_i	771	46	3	6	1	0	1	828
CIGP	760.94	56.65	8.43	1.57	0.33	0.07	0.02	828

Table 3: *Japanese labor force survey data (Sai and Takemura, 2000)*

[8] Hoshino, N. (2001). Applying Pitman's sampling formula to microdata disclosure risk assessment. *Journal of Official Statistics*, **17**, 499-520.

[9] Ismail, M.E.H. (1977). Integral representations and complete monotonicity of various quotients of Bessel functions. *Canadian Journal of Mathematics*, **29**, 1198–1207.

[10] Johnson, N.L., Kotz, S. and Balakrishnan, N. (1997). *Discrete Multivariate Distributions*, Wiley, New York.

[11] Jørgensen, B. (1982). *Statistical Properties of the Generalized Inverse Gaussian Distribution.* Lecture Notes in Statistics 9, Springer, New York.

[12] Neyman, J. (1939). On a new class of "contagious" distributions, applicable in entomology and bacteriology. *Annals of Mathematical Statistics*, **10**, 35–57.

[13] Reid, D.D. (1981). The Poisson lognormal distribution and its use as a model of plankton aggregation. *Statistical Distributions in Scientific Work*, C. Taillie, G.P. Patil and B. Baldessari Ed., **6**, Proceedings of the NATO Advanced Study Institute, 303–316, D. Reidel Publishing Company, Dordrecht.

[14] Sai, S. and Takemura, A. (2000). Some Models for Merging Groups in Microdata. *Japanese Journal of Applied Statistics*, **29**, 63–82 (in Japanese).

[15] Sankaran, M. (1968). Mixtures by the inverse Gaussian distribution. *Sankhyā*, **30**B, 455–458.

[16] Särndal, C.E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling.* Springer, New York.

[17] Seshadri, V. (1999). *The Inverse Gaussian Distribution.* Springer, New York.

[18] Sibuya, M. (1993). A random clustering process. *Annals of Institute of Statistical Mathematics*, **45**, 459–465.

[19] Sichel, H.S. (1971). On a family of discrete distributions particularly suited to represent long-tailed frequency data. *Proceedings of the Third Symposium on Mathematical Statistics (N.F. Laubscher, ed.)*, 51–97, S.A. C.S.I.R., Pretoria.

[20] Sichel, H.S. (1973). The density and size distribution of diamonds. *Bull. Int. Statist. Inst.*, **45**, 420–427.

[21] Sichel, H.S. (1982). Asymptotic efficiencies of three methods of estimation for the inverse Gaussian-Poisson distribution. *Biometrika*, **69**, 467–472.

[22] Sichel, H.S. (1997). Modelling species-abundance frequencies and species-individual functions with the generalized inverse Gaussian-Poisson distribution. *South African Statistical Journal*, **31**, 13–37.

[23] Stein, G.Z., Zucchini, W. and Juritz, J.M. (1987). Parameter Estimation for the Sichel distribution and its multivariate extension. *Journal of the American Statistical Association*, **82**, 938–944.

[24] Takemura, A. (1999). Some superpopulation models for estimating the number of population uniques. *Statistical data protection - Proceedings of the conference, Lisbon, 25 to 27 March 1998 - 1999 edition*, 45–58, Office for Official Publications of the European Communities, Luxembourg.

[25] Watson, G.N. (1944). *A Treatise on the Theory of Bessel Functions.* 2nd ed., University Press, Cambridge.

[26] Williams, C.B. (1956). Studies in the history of probability and statistics: IV. A note on an early statistical study of literary style. *Biometrika*, **43**, 248–256.

[27] Williams, C.B. (1964). *Patterns in the Balance of Nature.* Academic Press, London.

[28] Willenborg, L. and de Waal, T. (1996). *Statistical Disclosure Control in Practice* , Lecture Notes in Statistics 111, Springer, New York.