J. Japan Statist. Soc.
Vol. 33 No. 1 2003 105–117

RANDOM CLUSTERING BASED ON THE CONDITIONAL INVERSE GAUSSIAN-POISSON DISTRIBUTION

Nobuaki Hoshino*

The present article describes a Conditional Inverse Gaussian-Poisson (CIGP) distribution, obtained by conditioning an inverse Gaussian-Poisson population model on its total frequency. This CIGP distribution is equivalent to random partitioning of positive integers, with the possibility for a number of applications in statistical ecology, linguistics and statistical disclosure control to name a few. After showing the marginal moments of the distribution, parameter estimation is discussed. Fitting the CIGP distribution to some typical data sets demonstrates its applicability.

Key words and phrases: Disclosure risk, frequencies of frequencies, size index, species abundance, superpopulation.

1. Introduction

The observation of various kinds of populations includes many instances where the population consists of diverse groups with properties difficult to formulate. To comprehend this complex nature populations, it is often useful to focus upon the frequency structure of the population. This is a classical theme in statistics, dating back to e.g. Neyman (1939). We will later discuss more examples in which counting the size of groups in a population is of great importance. Models have been used to describe count data, and the present article proposes a new model of this type.

The organization of the present article is as follows. Section 1.1 describes the background. Section 1.2 derives the proposed model. Section 2 presents some theoretical results on this model. Section 3 discusses the parameter estimation of the model. Finally, Section 4 gives concluding remarks arising from the three application results.

1.1. Modeling count data

We consider a population of size N consisting of J cells (groups, species). In the following $F_j, j = 1, \ldots, J$, denotes the number of individuals (size or frequency) in the j-th cell. By definition, $N = \sum_{j=1}^{J} F_j$. The number of cells of size i is denoted by S_i. More specifically,

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

Received May 7, 2002. Revised September 17, 2002. Accepted March 13, 2003.

*Faculty of Economics, Kanazawa University, Kakuma-machi, Kanazawa-shi, Ishikawa, Japan. E-mail: hoshino@kenroku.kanazawa-u.ac.jp

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 the statistical literature, $(S_0, S_1, \dots)$ is called size indices (Sibuya (1993)) or frequencies of frequencies (Good (1953)). It also corresponds to the concept of equivalence class (Greenberg and Zayatz (1992)).

Obviously

$$\sum_{i=0}^{\infty} S_i = J, \qquad \sum_{i=1}^{\infty} i \cdot S_i = N, \tag{1.1}$$

where S_i is a nonnegative integer. The point to note is that J includes empty cells, which may represent unseen or extinct species.

For example, consider that a population is composed of J species, and F_j is the number of the j-th species in stochastic abundance models. See Engen (1978) for the context of statistical ecology. In addition to these examples, there are myriads of examples in linguistics, where a writer is deemed to use a vocabulary of J words, and each F_j expresses the frequency of the usage of the j-th word in the writer's text. The recent book by Baayen (2001) surveys developments on this field to the present.

In some of the applications described, the interest lies in knowing the population structure associated with size indices. In linguistics, S_1 is the number of *hapax legomena*, which are the words mentioned once only. In the area of disseminating microdata, an individual that is unique in a population is expediently considered to be identifiable. Thus S_1, the number of "population uniques", is a typical index of the risk of privacy invasion. See Willenborg and de Waal (1996, 2000) for the context of statistical disclosure control. Another example occurs in database merging. When databases overlap, an individual can appear more than once, where S_i is the number of i-fold appearances.

For practical purposes, it is necessary to estimate size indices based on samples. Engen (1978, Section 2.3) derived the unique unbiased estimator of S_i under simple random sampling without replacement. However, the variance of the estimator is impractically large. An estimator that has smaller variance requires more information about a population; the insufficiency can be compensated by assuming a distribution of size indices or a superpopulation. Following the empirical Bayes method, the parameter values of a superpopulation are estimated from samples, and $\mathrm{E}(S_i)$ under those estimates is the estimate of S_i. Therefore the marginal moments of size indices are especially important.

The major way to construct a superpopulation for count data is to regard F_j as an independent and identical mixed Poisson distribution. Then the population size N is a random variable, but conditioning such a model on N results in a distribution of a fixed number of individuals, which parallels a contingency table. It is equivalent also to random partitioning of the positive integers. See Hoshino

(2001) for a brief survey on existing models. The problem is that only a few models of this type are known to be manageable because of the combinatorics involved, and hence developing a new model of fixed size is of great importance.

If the distribution of F_j is closed under convolution, it is easy to derive the distribution of N. Then the construction of a conditional model $\mathrm{P}(F_1, F_2, \dots, F_J|N)$ becomes logically straightforward. Among distributions that are closed under convolution, the present article selects the inverse Gaussian-Poisson mixture proposed by Holla (1966) and investigates its conditional model.

The inverse Gaussian distribution is well reviewed by Seshadri (1993, 1999). In particular, Seshadri (1999) devotes its Section 7.1 to the Poisson mixture of the inverse Gaussian distribution. It is a special case of the generalized inverse Gaussian-Poisson mixture proposed by Sichel (1971), which is, however, less tractable. See Jørgensen (1982) for the generalized inverse Gaussian distribution. Concerning the (generalized) inverse Gaussian-Poisson mixture, there is 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. Because the inverse Gaussian-Poisson mixture has been used to describe populations, our approach seems to be promising for various applications.

1.2. The derivation of the Conditional Inverse Gaussian-Poisson distribution

The density of the inverse Gaussian (IG) distribution is for $0 < \theta \le 1, \alpha > 0$,

$$f(\lambda;\alpha,\theta) = \frac{(2\sqrt{1-\theta}/(\alpha\theta))^{(-1)/2}}{2K_{-1/2}(\alpha\sqrt{1-\theta})}\lambda^{-3/2}\exp\left(-\left(\frac{1}{\theta}-1\right)\lambda - \frac{\alpha^2\theta}{4\lambda}\right), \quad \lambda > 0, \tag{1.2}$$

where

$$K_{-1/2}(\alpha\sqrt{1-\theta}) = \sqrt{\frac{\pi}{2\alpha\sqrt{1-\theta}}}\exp(-\alpha\sqrt{1-\theta})$$

is the modified Bessel function of the third kind of order $-1/2$. It is noteworthy that the following discussion allows θ to be unity, where (1.2) reduces to the density of a stable distribution with exponent $1/2$. The reduced form is also called reciprocal gamma, Pearson type 5 or inverted gamma. In modeling a population, the IG distribution is important as a substitute for the log-normal distribution, which is a representative heavy-tailed distribution. The main difference is that the mixed distribution with the Poisson distribution is analytically tractable in the case of the IG distribution.

Suppose that a random variable Y is distributed as the Poisson distribution with mean λ, and let λ have density (1.2). Then the distribution of Y is

$$\mathrm{P}(Y = y;\alpha,\theta) = \sqrt{\frac{2\alpha}{\pi}}\exp(\alpha\sqrt{1-\theta})\frac{(\alpha\theta/2)^y}{y!}K_{y-1/2}(\alpha), \quad y = 0, 1, 2, \dots, \tag{1.3}$$

where $0 < \theta \leq 1, \alpha > 0$. When $\theta = 1$, (1.3) reduces to a special case of the discrete stable distribution proposed by Steutel and van Harn (1979). Sichel (1982) claimed that θ controls the upper tail of the distribution and α describes the lower. We adopt Sichel's parameterization because conditioning will remove the power parameter θ of (1.3). See Seshadri (1999, Chap. 7.1) for other kinds of parameterizations. The present article refers to (1.3) as the inverse Gaussian-Poisson distribution and denotes it by $IGP(\alpha, \theta)$.

For convenience, we cite two useful formulae of the modified Bessel function of the third kind, whose argument takes only real and positive value in the present article. First, from e.g. Watson (1944, Section 3.71),

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

Second, as shown by Ismail (1977),

$$K_\gamma(\alpha) \approx 2^\gamma \gamma^\gamma \exp(-\gamma) \alpha^{-\gamma} \sqrt{\frac{\pi}{2\gamma}} \tag{1.4}$$

when γ is large. Equation (1.4) is useful because the computation of the modified Bessel function of the third kind requires more and more resources as $\gamma \to \infty$.

When $F_j, j = 1, 2, \dots, J$, are independently and identically distributed as $IGP(\alpha, \theta)$,

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

or

$$\mathrm{P}(S_0, S_1, \dots) = J! \prod_{i=0}^{\infty} \left\{ \sqrt{\frac{2\alpha}{\pi}} \exp(\alpha\sqrt{1-\theta}) \frac{(\alpha\theta/2)^i}{i!} K_{i-1/2}(\alpha) \right\}^{S_i} \frac{1}{S_i!}. \tag{1.5}$$

The conditional distribution of (1.5) given N is the Conditional Inverse Gaussian-Poisson (CIGP) distribution, which we will define in the next section. As mentioned before, it is easy to derive the exact distribution of the population size N under the IGP model (1.5). The probability generating function of (1.3) is

$$G(z) = \exp(\alpha(\sqrt{1-\theta} - \sqrt{1-z\theta})), \tag{1.6}$$

which can be verified using the fact that $\sum_{y=0}^{\infty} \mathrm{P}(y) = 1$. Consequently, the sum of J random variables that are independently identically distributed as $IGP(\alpha, \theta)$ is distributed as $IGP(J\alpha, \theta)$. That is,

$$\mathrm{P}(N) = \sqrt{\frac{2J\alpha}{\pi}} \exp(J\alpha\sqrt{1-\theta}) \frac{(J\alpha\theta/2)^N}{N!} K_{N-1/2}(J\alpha), \quad N = 0, 1, 2, \dots , \tag{1.7}$$

where $0 < \theta \leq 1,\ \alpha > 0$.

2. On the property of the CIGP distribution

The main result of this section is to show the factorial moments of size indices under the CIGP distribution, whose sampling distribution will also be discussed.

Dividing (1.5) by (1.7), the CIGP distribution is defined for $\alpha > 0$ as

$$\begin{aligned} &\mathrm{P}_J(S_0, \dots, S_N|N) \\ &= \left(\frac{2\alpha}{\pi}\right)^{(J-1)/2} \frac{J!N!}{J^{N+1/2}K_{N-1/2}(J\alpha)} \prod_{i=0}^{N} \left\{\frac{K_{i-1/2}(\alpha)}{i!}\right\}^{S_i} \frac{1}{S_i!}, \end{aligned} \tag{2.1}$$

where the argument satisfies (1.1). In contrast to $IGP(\alpha, \theta)$, the CIGP distribution owns only one parameter. The conditioning has removed θ, which makes sense because N is sufficient for the power parameter of a power series distribution. In the following $CIGP(\alpha)$ indicates (2.1).

First of all, $CIGP(\alpha)$ is an analogue of the Dirichlet-multinomial mixture proposed by Mosimann (1962). Rewriting (2.1) as $\mathrm{P}(F_1, F_2, \dots, F_J|N)$ may suggest this point more clearly. Sibuya *et al.* (1964) noted that the conditional model of the gamma-Poisson mixture (=negative binomial distribution) given N equals the Dirichlet-multinomial mixture or multivariate negative hypergeometric distributions. Hoshino (2002a) investigates parallel relationships of the CIGP distribution to those of the Dirichlet-multinomial mixture, based on Hoshino and Takemura (1998)'s discussion. Both the negative binomial distribution and the IGP distribution belong to the class of infinitely divisible distributions on the nonnegative integers, and Hoshino (2002b) generalizes these relationships over this class.

We next show the expectations of size indices. The following theorem reveals some combinatorial facts as by-products; see Appendix.

THEOREM 1. *Suppose that size indices are distributed as* (2.1). *Then the factorial moments are*

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

where $r = \sum_{j=1}^{N} r_j \le J$, $R = \sum_{j=1}^{N} j r_j \le N$, *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

$$\mathcal{S}(N) = \left\{ \boldsymbol{S} = (S_1, \dots, S_N) \middle| \sum_{i \ge 1} i S_i = N \right\}.$$

Table 1. Expectations of size indices under $N = 1000, J = 10000$.

α	$\mathrm{E}(S_1)$	$\mathrm{E}(S_2)$	$\mathrm{E}(S_3)$	$\mathrm{E}(S_4)$	$\mathrm{E}(S_5)$
0.1	391.01	89.14	37.05	19.20	11.13
0.5	749.38	92.10	15.92	3.31	0.76
1.0	823.11	74.44	7.85	0.94	0.12
10.0	896.38	48.77	1.94	0.06	0.00
50.0	903.21	45.93	1.59	0.04	0.00

Then

$$
\begin{aligned}
&\mathrm{E}_J(S_j|N) \\
&\quad = \sum_{\boldsymbol{S}\in\mathcal{S}(N)} S_j \, \mathrm{P}_J(S_0, \dots, S_N|N) \\
&\quad = \sqrt{\frac{2\alpha}{\pi}} \left(\frac{K_{j-1/2}(\alpha)}{j!} \right) \frac{N! K_{N-j-1/2}((J-1)\alpha)(J-1)^{N-j+1/2}}{(N-j)! J^{N-1/2} K_{N-1/2}(J\alpha)} \\
&\qquad \times \sum_{\boldsymbol{S}\in\mathcal{S}(N)} \mathrm{P}_{J-1}(S_0, \dots, S_{j-1}, S_j - 1, S_{j+1}, \dots, S_N|N-j) \, I(S_j \geq 1).
\end{aligned} \tag{2.2}
$$

Since

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

we obtain

$$
\mathrm{E}_J(S_j|N) = \sqrt{\frac{2\alpha}{\pi}} \left(\frac{K_{j-1/2}(\alpha)}{j!} \right) \frac{N! K_{N-j-1/2}((J-1)\alpha)(J-1)^{N-j+1/2}}{(N-j)! J^{N-1/2} K_{N-1/2}(J\alpha)} \tag{2.3}
$$

from (2.2). We have only evaluated $E(S_j|N)$, but $\mathrm{E}\left(\prod_{j=1}^N S_j^{(r_j)}|N\right)$ can be evaluated with the same argument. □

Table 1 shows the expectations of size indices with parameter values $\alpha = 0.1, 0.5, 1, 10, 50$ under $N = 1000$ and $J = 10000$. It is observable that $\mathrm{E}(S_i)$ is decreasing with respect to i; this is the pattern frequently found in applications. The difference between the size indices of $\alpha = 10$ and $\alpha = 50$ appears very small compared to that of $\alpha = 0.1$ and $\alpha = 0.5$.

When N is large, it is convenient to use an approximation to the expectation of a size index. The following proposition shows asymptotic expressions of $E(S_1|N)$, which is of great significance in applications. Equation (2.4) is an immediate consequence of (1.4) and (2.3). The second expression (2.5) was suggested by a referee. Section 4 exemplifies applications of these formulae.

PROPOSITION 1. *Suppose that size indices are distributed according to* (2.1). *Then, as* $N \to \infty$,

$$\mathrm{E}(S_1|N) \approx \exp(1-\alpha)\frac{N\alpha(J-1)}{2}\frac{(N-3/2)^{N-2}}{(N-1/2)^{N-1}} \tag{2.4}$$

or

$$\mathrm{E}(S_1|N) = \exp(-\alpha)\alpha(J-1)/2 + O(N^{-1}). \tag{2.5}$$

Under $CIGP(\alpha)$, we now discuss the unconditional distribution of sample size indices, which is called "sampling distribution". Let us denote the sample size by n; the sample size indices are similarly defined and denoted by $(s_0, s_1, \dots)$. Because the CIGP distribution does not depend on the label of each individual, Lemma 1 of Takemura (1999) ensures that the sampling distribution is the result of substituting n for N of the population distribution. In other words, the distribution of n samples directly drawn from the infinite population is the same as that of n samples drawn from the finite population of size N, which is in turn drawn from the infinite population.

THEOREM 2. *Suppose that the distribution of population size indices is* (2.1) *and denoted by* $\mathrm{P}(S_0, \dots, S_N|N;\alpha)$. *Let* n *samples be drawn with simple random sampling without replacement from the population. Then the sample size indices are distributed as* $\mathrm{P}(s_0, \dots, s_n|n;\alpha)$.

3. Parameter estimation

This section treats the estimation of α from samples that are distributed as $CIGP(\alpha)$. The subjects are the Maximum Likelihood (ML) estimator and two approximate estimators.

3.1. Maximum likelihood estimation

Given Theorem 2, the log likelihood of sample size indices is

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

In the following

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

As noted in e.g. Seshadri (1999, p. 125),

$$\frac{\partial \log K_\gamma(\alpha)}{\partial \alpha} = -R_\gamma(\alpha) + \frac{\gamma}{\alpha}. \tag{3.1}$$

Using (3.1), we construct the ML estimator: the solution of $dL/d\alpha = 0$. From (1.1),

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

The ML estimation obviously requires numerical evaluation. The Newton-Raphson method is available based on the second derivative:

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

The starting value of such an iterative procedure may be given by the estimators that will be discussed in Section 3.2.

3.2. Approximate estimation

The expectation of a sample size index is the result of substituting n for N in (2.3). Based on this fact, the method of moments can apply, but it needs numerical evaluation. Instead we derive easy-to-calculate approximate estimators from existing estimators of the IGP distribution, though the property of these approximations is not clear. See Section 4 for an empirical comparison of these.

The first to consider is an approximate moment estimator. If Y is a random variable that is subject to $IGP(\alpha, \theta)$,

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

and

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

See the probability generating function (1.6). The sample average n/J substitutes for $\mathrm{E}(Y)$, and the sample variance

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

substitutes for $\mathrm{V}(Y)$. The solution to these simultaneous equations results in an approximate estimator:

$$\tilde{\alpha} = \frac{n\sqrt{n(2Jv-n)}}{J(Jv-n)}. \tag{3.2}$$

Table 2. Frequency distribution of *Oithona similis* nauplii (Barnes and Marshall, 1951).

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)

Sichel (1982) calculated asymptotic efficiencies for the joint estimation of α and θ for the method of moments. However, the moment estimators were inefficient when parameter values are typical in situations where $IGP(\alpha, \theta)$ is fitted to real data. Hence (3.2) may not be efficient in practice.

The second to consider is an approximation based on estimators proposed by Sichel (1973). These estimators were efficient in the typical range of values according to Sichel's calculation. It leads to

$$\bar{\alpha} = -\frac{1}{2}(\log s_0 - \log J)\left(1 + \frac{n/J}{n/J + \log s_0 - \log J}\right) \tag{3.3}$$

in our setting.

4. Application results and a conclusion

This section applies $CIGP(\alpha)$ to real data. The IGP distribution is known to fit the claim frequency of insurance extremely well; see Willmot (1987) for example. However, we dare to select three data sets from other areas. After some discussions, the present article will conclude.

Table 2 Reid (1981) fitted Log-Normal-Poisson mixture (LNP), Gamma-Poisson mixture (GP) and the Neyman type A (NY) distribution to plankton data of $n = 232$ and $J = 120$ provided by Barnes and Marshall (1951). The CIGP distribution is also fitted to the same data; the ML estimate appears to be $\hat{\alpha} = 10.35$, and the fitted values of size indices are the expectations under $\hat{\alpha}$. The approximate moment estimate by (3.2) is $\tilde{\alpha} = 4.49$, and another estimate by (3.3) is $\bar{\alpha} = 6.50$; these are not close to the ML estimate. The χ^2 value of the CIGP distribution is provided just for comparison and not suitable for the test of fit.

Table 3 Stein *et al.* (1987) fitted the IGP distribution to William (1964)'s lice data, where $n = 7442$ and $J = 1083$. The CIGP distribution can attain a comparable fit to that of the IGP distribution, though these fits are not good in this case. The approximate estimates are $\tilde{\alpha} = 1.069$ and $\bar{\alpha} = 0.579$.

Table 3. Frequency distribution of lice (Williams, 1964).

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 ML		0.645	0.644
$\hat{\theta}$ by ML		0.998	

Table 4. Japanese Labor Force Survey data (Sai and Takemura, 2000).

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

Sichel (1982) remarked that the correlation between the ML estimators of $IGP(\alpha, \theta)$ is generally substantial in the useful range of values. Because it causes numerical instability, Stein et al. (1987) proposed a reparameterization. In this context, the CIGP distribution is valuable since it involves no numerical instability at a similar fit.

Table 4 Sai and Takemura (2000) anonymized Japanese Labor Force Survey data collected in December 1997 and calculated their size indices. We apply the CIGP distribution to one of the sets, which is interesting because there seem-

ingly exists no application of the IGP distribution to the inference of population uniques. In this case, $J = 5.644 \times 10^{12}$ and $n = 908$. 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 number of nonempty groups is denoted by $u = \sum_{i \geq 1} s_i$. The population size N equals 1.028 million, where the exact value of $\mathrm{E}(S_1|N)$ amounts to 2553.05 under the ML estimate. Its approximate values by (2.4) and (2.5) are 2553.07 and 2553.06 respectively. Sai and Takemura's discussion was equivalent to the evaluation that $S_1 \doteq 3368$.

Concluding remarks Fitting the CIGP distribution needs the information of s_0, but many data in practice have no information of s_0. In such a case, we can use the limiting distribution of $CIGP(\alpha)$ derived in Hoshino (2002a). The estimate by (3.3) has been closer to the ML estimate than that of (3.2) in our experiments, which suggests that $\bar{\alpha}$ may be better than $\tilde{\alpha}$ on real data. The CIGP distribution can surely be a model of count data. In particular, it has merits in being free from the numerical instability that occurs in the ML estimation of the IGP parameters.

Appendix

An anonymous referee suggested the following multiple formulae of the modified Bessel function of the third kind of a half-odd integer order. Let J and N be positive integers. When $\alpha > 0$,

$$
\begin{aligned}
& K_{N-1/2}(J\alpha) \\
(\mathrm{A.1}) \quad & = \sqrt{\frac{2\alpha}{\pi}} \left(\frac{J-1}{J} \right)^{N+1/2} \\
& \quad \times \sum_{j=0}^{N} \binom{N}{j} K_{j-1/2}(\alpha) K_{N-j-1/2}((J-1)\alpha)(J-1)^{-j} \\
(\mathrm{A.2}) \quad & = \sqrt{\frac{2\alpha}{\pi}} \left(\frac{J-1}{J} \right)^{N-1/2} \\
& \quad \times \sum_{j=1}^{N} \binom{N-1}{j-1} K_{j-1/2}(\alpha) K_{N-j-1/2}((J-1)\alpha)(J-1)^{-j+1} \\
(\mathrm{A.3}) \quad & = \sum_{(S_0, S_1, \ldots, S_N) \in \mathcal{S}^*} \left(\frac{2\alpha}{\pi} \right)^{(J-1)/2} \frac{J!N!}{J^{N+1/2}} \prod_{j=0}^{N} \left\{ \frac{K_{j-1/2}(\alpha)}{j!} \right\}^{S_j} \frac{1}{S_j!},
\end{aligned}
$$

where $\mathcal{S}^* = \{(S_0, S_1, \ldots, S_N) \mid \sum_{j=1}^{N} jS_j = N,\ \sum_{j=0}^{N} S_j = J\ \}$. Using (2.3), we can show that $\sum_{j=0}^{N} \mathrm{E}(S_j) = J$ is equivalent to (A.1) and $\sum_{j=1}^{N} j\mathrm{E}(S_j) = N$ is equivalent to (A.2). Equation (A.3) holds because the sum of the probability is unity.

Acknowledgements

This manuscript was mainly prepared while the author was visiting Carnegie Mellon University. The author was financed by the Japanese Ministry of Education, Culture, Sports, Science and Technology. Prof. S. Fienberg, Prof. M. Sibuya, Prof. A. Takemura and anonymous referees provided valuable comments. The author would like to express sincere thanks to them.

References

Baayen, R. H. (2001). *Word Frequency Distributions*, Kluwer, Dordrecht.

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.

Engen, S. (1978). *Stochastic Abundance Models*, Chapman and Hall, London.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters, *Biometrika*, **40**, 237–264.

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

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

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

Hoshino, N. (2002a). On limiting random partition structure derived from the conditional inverse Gaussian-Poisson distribution, Technical Report CMU-CALD-02-100, School of Computer Science, Carnegie Mellon University.

Hoshino, N. (2002b). Engen's extended negative binomial model revisited, *Discussion Paper No. 2002-1*, Faculty of Economics, Kanazawa University.

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.

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

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

Mosimann, J. E. (1962). On the compound multinomial distribution, the multivariate β-distribution and correlations among proportions, *Biometrika*, **49**, 65–82.

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

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

Sai, S. and Takemura, A. (2000). Some models for merging groups in microdata, *Japanese Journal of Applied Statistics*, **29**, 63–82 (in Japanese).

Seshadri, V. (1993). *The Inverse Gaussian Distribution: A Case Study in Exponential Families*, Clarendon Press, Oxford.

Seshadri, V. (1999). *The Inverse Gaussian Distribution*, Lecture Notes in Statistics 137, Springer, New York.

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

Sibuya, M., Yoshimura, M. and Shimizu, R. (1964). Negative multinomial distribution, *Annals of Institute of Statistical Mathematics*, **16**, 409–426.

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*, (ed. N. F. Laubscher), 51–97, S.A. C.S.I.R., Pretoria.

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

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

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.

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.

Steutel, F. W. and van Harn, K. (1979). Discrete analogues of self-decomposability and stability, *Annals of Probability*, **7**, 893–899.

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.

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

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

Willenborg, L. and de Waal, T. (2000). *Elements of Statistical Disclosure Control*, Lecture Notes in Statistics 155, Springer, New York.

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

Willmot, G. E. (1987). The Poisson-inverse Gaussian distribution as an alternative to the negative binomial, *Scandinavian Actuarial Journal*, 113–127.