# On Limiting Random Partition Structure Derived from the Conditional Inverse Gaussian-Poisson Distribution

Nobuaki Hoshino[1]

February 2002
CMU-CALD-02-100

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213



**Abstract**

In the present article, we derive a new multivariate distribution that belongs to an exponential family through a limiting argument over the conditional inverse Gaussian-Poisson distribution proposed by Hoshino (2001b). The derived distribution can be used as a model of random partitioning of positive integers, which is relevant to applications in many fields such as statistical ecology, linguistics and statistical disclosure control to name a few. We clarify some properties of this distribution that are important in applications.

[1]Visiting faculty from 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 and Prof. Takemura's useful comments on the subject.

**Keywords:** Random clustering, Species abundance, Superpopulation, Disclosure risk

# 1 Introduction

Population modeling is a classical theme in statistics. In regard to a population composed of heterogeneous groups, a research often focuses upon the structure of frequencies of individuals. Therefore statisticians have been developing a number of population models that can describe frequency structures. For instance, Fisher et al. (1943) enumerated species in a population of Malayan butterflies and summarized the information with the logarithmic series distribution. This study is followed by many a scholar, and they constitute the field of statistical ecology or stochastic abundance models; see Engen (1978) for more detail. We can find analogous examples in various other fields: See Baayen (2001) for the context in linguistics and Hoshino (2001a) for a succinct survey in statistical disclosure control.

Mixed Poisson distributions play a central role in population modeling. The quintessence of these mixtures is gamma-Poisson, which equals negative binomial. Some other models relate to a gamma-Poisson population model: We can derive the Dirichlet-multinomial model (Takemura (1999)) from the gamma-Poisson model by conditioning the population size, and we can obtain the logarithmic series model (Anscombe (1950)) from the gamma-Poisson model through a limiting argument that resembles the law of small numbers. The Ewens distribution (see Chap. 41 of Johnson et al. (1997)) can be obtained from the logarithmic series model by conditioning the population size or from the Dirichlet-multinomial model through a limiting argument that is the same as one used to derive the logarithmic series model from the gamma-Poisson model. These relationships are summarized in Figure 1 from Hoshino and Takemura (1998), who discussed these facts in detail .

Recently Hoshino (2001b) proposed the Conditional Inverse Gaussian-Poisson (CIGP) distribution, which was derived from an inverse Gaussian-Poisson population model by conditioning the population size. A limit of the inverse Gaussian-Poisson mixture is the reciprocal gamma-Poisson mixture. The CIGP distribution hence corresponds to the Dirichlet-multinomial model in a sense; it is natural to investigate distributions that correspond to the Ewens distribution and the logarithmic series model.

In the present article we derive limiting distributions of the inverse Gaussian-Poisson model and the CIGP distribution with the law-of-small-numbers-like argument. In particular our discussion will concentrate upon the property of the limiting distribution of the CIGP distribution; the derived distribution seems new and treatable. Models that are conditioned on the population size are equivalent to the random partitioning of positive integers, which often involves complicated combinatorics. In consequence only a few such models are known as easily tractable. It is thus valuable to develop models that are conditioned on the population size, in order to handle diverse populations.

The organization of the present article is as follows. In Section 2 we provide notation and definitions of existing models. In Section 3 we derive the limiting distribution of the CIGP distribution and clarify its properties. In Section 4 we discuss the parameter estimation of the proposed distribution. In Section 5 we apply the proposed distribution to a typical data set for exemplifying the usefulness of the distribution and conclude.

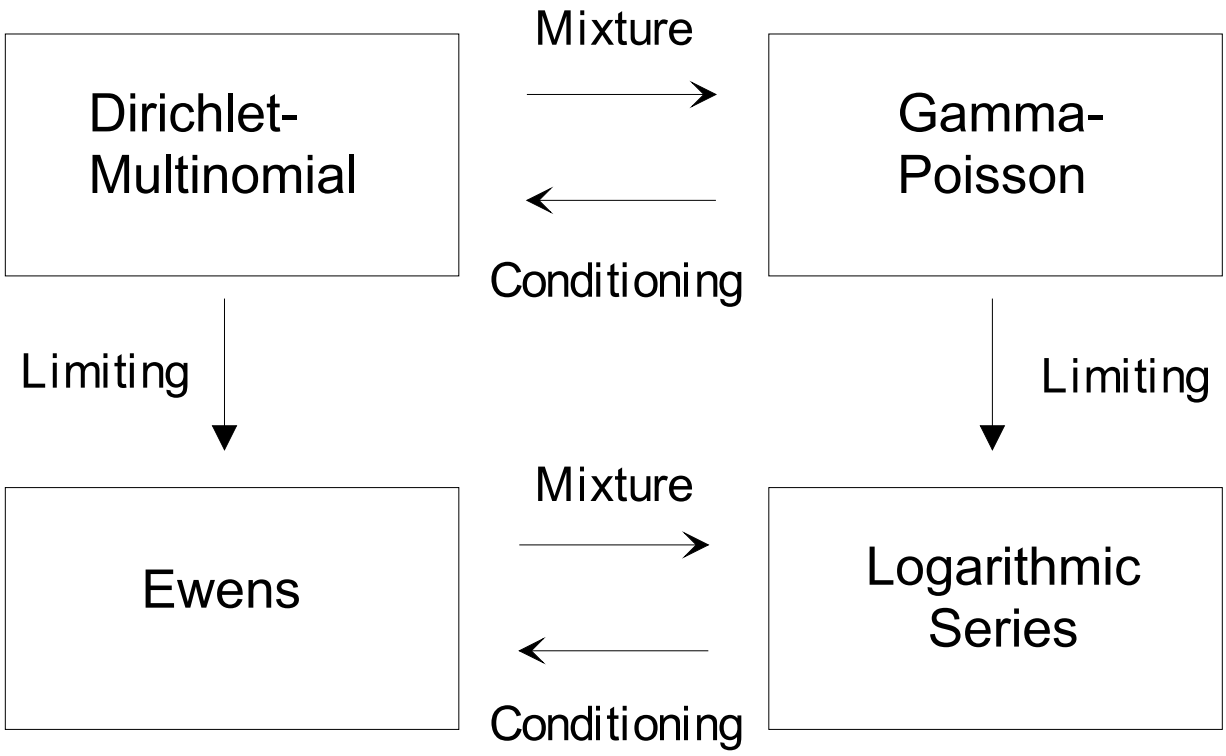


Figure 1: *Relationships among gamma-Poisson-related models*

## 2 Backgrounds

Consider a population of size $N$ consisting of $J$ cells (groups, species, words) with the size (frequency) $F_j, j = 1, \ldots, J$, $N = \sum_{j=1}^{J} F_j$. 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)). A linguist may be interested in the number of nonce words or *hapax legomena* in a text, which corresponds to $S_1$. A statistical agency may also be interested in $S_1$, which is called "population uniques" and is a major risk-index of disclosure; see Willenborg and de Waal (1996, 2000).

Obviously

$$\sum_{i=0}^{\infty} S_i = J,$$

$$\sum_{i=1}^{\infty} i \cdot S_i = N. \tag{1}$$

Note that $J$ is the total number of cells including the number of the empty cells $S_0$. Empty cells may correspond to unseen or extinct species. In the following we denote the number of non-empty cells by

$$U = \sum_{i=1}^{\infty} S_i = J - S_0. \tag{2}$$

When $N$ is given, we write in particular

$$U_N = \sum_{i=1}^{N} S_i,$$

since $S_i$ equals zero for $i \geq N + 1$. In the stochastic abundance model, $U_N$ corresponds to the total number of species within a population of size $N$; Bunge and Fitzpatrick (1993) surveyed this problem of estimating the number of species.

In order to consider sampling distributions explicitly, we denote the sample size by $n$ and sample size indices by $(s_0, s_1, \ldots)$, which are defined in the same way as those of the population. The number of non-empty cells of samples is

$$u = \sum_{i=1}^{\infty} s_i = J - s_0.$$

When $n$ is given, we use $u_n = \sum_{i=1}^{n} s_i$.

Let us assume that random variables $F_j, j = 1, \ldots, J$, are independently and identically distributed as the Inverse Gaussian-Poisson (IGP) distribution, which is well reviewed in Chap.

7.1 of Seshadri (1999). In terms of the size indices, the IGP population model is then expressed, for $0 < \theta < 1, \alpha > 0$, as

$$\mathrm{P}(S_0, S_1, \ldots) = J! \prod_{i=0}^{\infty} \{\sqrt{\frac{2\alpha}{\pi}} \exp(\alpha\sqrt{1-\theta}) \frac{(\alpha\theta/2)^i}{i!} K_{i-1/2}(\alpha)\}^{S_i} \frac{1}{S_i!}, \tag{3}$$

where $K_{i-1/2}(\cdot)$ is the modified Bessel function of the third kind of order $i - 1/2$. The limiting form of (3) as $\theta \to 1$ is the reciprocal gamma-Poisson model.

From

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

we obtain

$$K_{y-1/2}(\alpha) = \sqrt{\frac{\pi}{2\alpha}} \exp(-\alpha) (\sum_{i=0}^{y-1} \frac{(y-1+i)!}{(y-1-i)!i!} (2\alpha)^{-i}), \quad y = 1, 2, \ldots, \tag{4}$$

using the fact that $K_{-1/2}(\alpha) = K_{1/2}(\alpha)$ and the following recurrence formula:

$$K_{\gamma+1}(\alpha) = \frac{2\gamma}{\alpha} K_{\gamma}(\alpha) + K_{\gamma-1}(\alpha). \tag{5}$$

Consult Watson (1944) for the results on Bessel functions.

According to Hoshino (2001b), the population size $N$ under (3) is distributed as

$$\mathrm{P}(N) = \sqrt{\frac{2J\alpha}{\pi}} \exp(J\alpha\sqrt{1-\theta}) \frac{(J\alpha\theta/2)^N}{N!} K_{N-1/2}(J\alpha), \tag{6}$$

and the Conditional Inverse Gaussian-Poisson (CIGP) distribution is obtained by dividing (3) by (6). For $\alpha > 0$, the CIGP distribution was thus defined as

$$\mathrm{P}(S_0, \ldots, S_N) = (\frac{2\alpha}{\pi})^{\frac{J-1}{2}} \frac{J!N!}{J^{N+1/2} K_{N-1/2}(J\alpha)} \prod_{i=0}^{N} \{\frac{K_{i-1/2}(\alpha)}{i!}\}^{S_i} \frac{1}{S_i!}. \tag{7}$$

In Hoshino (2001b) the CIGP distribution (7) was fitted to three data sets. The maximum likelihood estimate of $\alpha$ and $J$ for these data sets were $(10.35, 120), (0.645, 1083)$ and $(9.047 \times 10^{-10}, 5.644 \times 10^{12})$, respectively. The last set concerns microdata disclosure risk assessment, and $J$ is usually very large in this field. Therefore it is realistic to consider the limiting case of $J \to \infty$ with $J\alpha$ fixed.

## 3 Main Results

This section clarifies limiting properties of the CIGP distribution (7). All the proofs of theorems in this section are provided in Appendix.

**Theorem 1** *Let $J\alpha = A$ $(> 0)$ be fixed and let $J \to \infty, \alpha \to 0$. Then the limiting distribution of the CIGP distribution* (7) *is*

$$\mathrm{P}(S_1, \ldots, S_N) = \sqrt{\frac{\pi}{2A}} \frac{N! \exp(-A)}{A^{N-U_N} K_{N-1/2}(A)} \prod_{i=1}^{N} \{\frac{(2i-3)!!}{i!}\}^{S_i} \frac{1}{S_i!}, \tag{8}$$

*where* $(-1)!! = 1, (2i-3)!! = (2i-3)(2i-5)\cdots 1$.

It is worthy of note that $S_0$ is no longer defined in (8) owing to $J$ being infinity. In applications we often have no information on groups of zero frequencies, which prevents us from applying models that depend on $S_0$ such as the CIGP distribution. We can apply (8) in this case and may be able to regard (8) as a zero-truncated version of the CIGP distribution; the discussion on zero-truncation will continue after Theorem 2.

We can construct (8) in another way. The relationships stated in Theorem 2 are analogous to those among the gamma-Poisson-related models; compare Figure 2 with Figure 1.

**Theorem 2** *Let $J\alpha = A$ $(> 0)$ be fixed and let $J \to \infty, \alpha \to 0$. Then the limiting distribution of the IGP population model* (3) *is*

$$\mathrm{P}(S_1, S_2, \ldots) = \exp(A(\sqrt{1-\theta} - 1)) \prod_{i=1}^{\infty} \frac{{\tau_i}^{S_i}}{S_i!} = \prod_{i=1}^{\infty} \frac{\exp(-\tau_i){\tau_i}^{S_i}}{S_i!}, \tag{9}$$

*where*

$$\tau_i = A(\frac{\theta}{2})^i \frac{(2i-3)!!}{i!} = \frac{A\theta^i}{2\sqrt{\pi}} \frac{\Gamma(i-1/2)}{\Gamma(i+1)}.$$

*Namely, each $S_i, i = 1, 2, \ldots$, is independently distributed as Poisson with mean $\tau_i$.*

*The conditional distribution of* (9) *given its population size $N$ is* (8)*, and conversely the mixture of* (8) *by the population size distribution* (6) *leads to* (9)*.*

As we have mentioned, Anscombe (1950)'s interpretation of Fisher et al. (1943) is as follows. The same limiting argument that we employed in Theorem 1 and 2 over a gamma-Poisson (negative binomial) population model leads to the logarithmic series model: For $i = 1, 2, \ldots$, each $S_i$ is independently distributed as Poisson with mean $\lambda_i$, where $\lambda_i$ is proportional to $c^i/i$.

On the other hand, Fisher et al. (1943) is widely recognized as having proposed the logarithmic series distribution:

$$\mathrm{P}(X = x) = \frac{1}{-\log(1-\theta)} \frac{\theta^x}{x}, \quad x = 1, 2, \ldots, \tag{10}$$

where $0 < \theta < 1$. The logarithmic series distribution is a limit of the zero-truncated negative binomial distribution (Sampford (1955)); see Chap. 7 of Johnson et al. (1993). Note that (10) has the same power series structure of the logarithmic series model (i.e. $\lambda_i \propto c^i/i,\ i = 1, 2, \ldots$).

We can see an analogous relationship between $\tau_i, i = 1, 2, \ldots$, of (9) and a limit of the zero-truncated IGP distribution:

$$\mathrm{P}(X = x) = \sqrt{\frac{2\alpha}{\pi}} \frac{\exp(\alpha)}{\exp(\alpha(1 - \sqrt{1-\theta})) - 1} \frac{(\alpha\theta/2)^x}{x!} K_{x-1/2}(\alpha), \quad x = 1, 2, \ldots, \tag{11}$$

which was used by Sichel (1975).

**Theorem 3** *The limiting distribution of the zero-truncated IGP distribution* (11) *as $\alpha \to 0$ is*

$$\mathrm{P}(X = x) = \frac{1}{1 - \sqrt{1-\theta}} \frac{\theta^x (2x-3)!!}{2^x x!}, \quad x = 1, 2, \ldots, \tag{12}$$

*where $0 < \theta < 1$.*

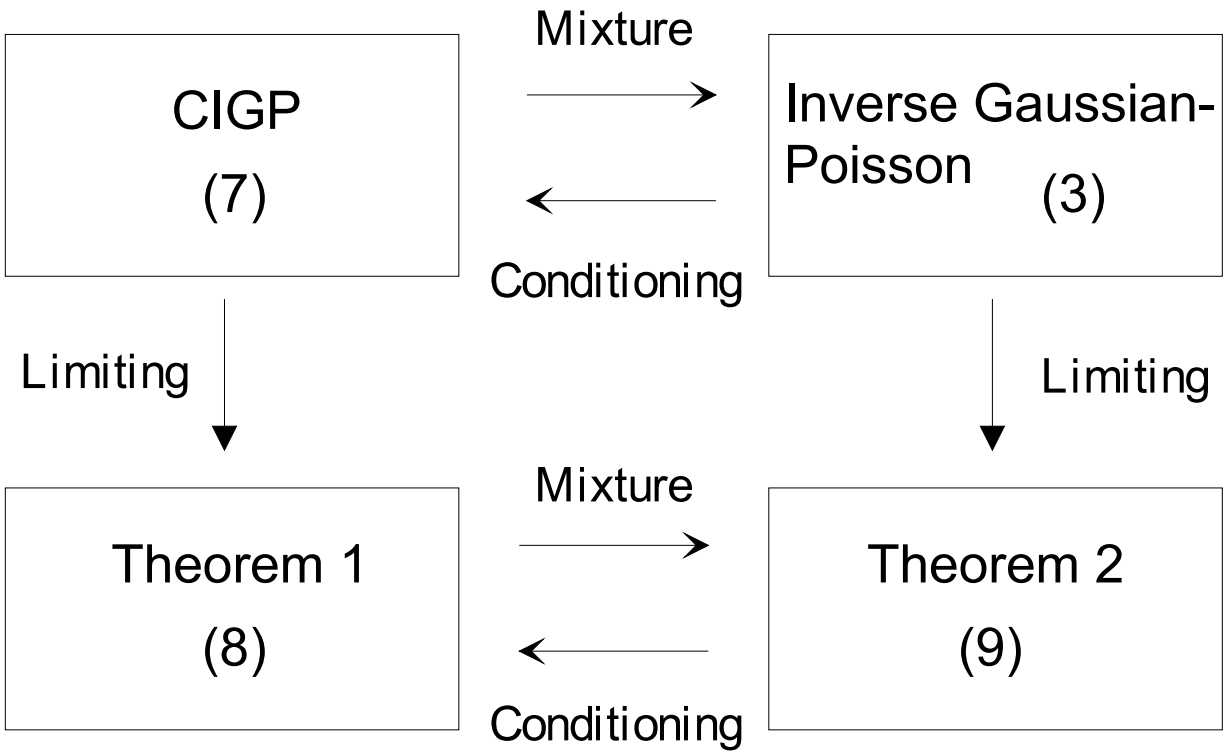


Figure 2: *Relationships among inverse-Gaussian-Poisson-related models*

Equation (12) is proportional to $\tau_i$ in (9) by putting $i = x$. The distribution (12) as well as the logarithmic series distribution belongs to the class of the power series distribution (Noack (1950)), whose properties are summarized in Johnson et al. (1993, p.70). Since equations (9) and (12) are free from the modified Bessel function, these distributions are convenient to manipulate.

Using (4), we can rewrite (8) as

$$\begin{aligned} &\mathrm{P}(S_1, \ldots, S_N) \\ &= \exp(U_N \log A + \log(\prod_{i=1}^{N} \{\frac{(2i-3)!!}{i!}\}^{S_i} \frac{1}{S_i!}) - \log(A^N \sum_{i=0}^{N-1} \frac{(N-1+i)!}{(N-1-i)!i!}(2A)^{-i}) + \log N!), \end{aligned}$$

which shows the following fact.

**Remark 1** *The distribution* (8) *belongs to an exponential family, and its sufficient statistics is* $U_N$.

A similar result to Remark 1 holds in the Ewens sampling formula; see Sibuya (1991). It should be remembered that the sufficient statistics $U_N$ is an important variate in applications. For instance, we can propose a new method of the estimation of the total number of species, based on the property of $U_N$.

We now derive the distribution of $U_N$.

**Theorem 4** *Suppose that size indices are distributed as* (8). *Then*

$$\mathrm{P}(U_N) = \sqrt{\frac{\pi}{2A}} \frac{\exp(-A)}{K_{N-1/2}(A)} (\frac{1}{2A})^{N-U_N} \frac{(2N-U_N-1)!}{(U_N-1)!(N-U_N)!}. \tag{13}$$

From (5) we have the the following recurrence formula on the distribution of $U_N$:

$$\mathrm{P}(U_{N+1} = v) = \frac{K_{N-1/2}(A)}{K_{N+1-1/2}(A)} \frac{2N-1}{A} \mathrm{P}(U_N = v) + \frac{K_{N-1-1/2}(A)}{K_{N+1-1/2}(A)} \mathrm{P}(U_{N-1} = v-2).$$

The Ewens sampling formula, as pointed out by Sibuya (1993), has the following urn model implication:

$$\begin{aligned} \mathrm{P}_{Ewens}(U_{N+1} = v) &= \mathrm{P}_{Ewens}(U_{N+1} = v | U_N = v) \mathrm{P}_{Ewens}(U_N = v) \\ &\quad + \mathrm{P}_{Ewens}(U_{N+1} = v | U_N = v-1) \mathrm{P}_{Ewens}(U_N = v-1), \end{aligned}$$

which provides profound understanding on the model; it is natural to seek an analogous implication of (8). However, the author has not found a good explanation.

Since $U_N = \sum_{i=1}^{N} S_i$, we can calculate the moments of $U_N$ using the moments of $S_i$. We hence evaluate the joint factorial moments of size indices, which are useful also for other applications than the estimation of the total number of species.

**Theorem 5** *Suppose that size indices are distributed as* (8). *Then the factorial moments are*

$$\mathrm{E}(\prod_{i=1}^{N} S_i^{(r_i)}) = \frac{K_{N-R-1/2}(A) A^{r-R} N!}{K_{N-1/2}(A)(N-R)!} \prod_{i=1}^{N} (\frac{(2i-3)!!}{i!})^{r_i}, \tag{14}$$

*where* $r = \sum_{i=1}^{N} r_i$, $R = \sum_{i=1}^{N} i r_i$ *and* $n^{(R)} = n(n-1)\cdots(n-R+1)$.

| $A$ | $\mathrm{E}(S_1)$ | $\mathrm{E}(S_2)$ | $\mathrm{E}(S_3)$ | $\mathrm{E}(S_4)$ | $\mathrm{E}(S_5)$ |
|---|---|---|---|---|---|
| 0.10 | 0.05 | 0.01 | 0.01 | 0.00 | 0.00 |
| 1.00 | 0.50 | 0.13 | 0.06 | 0.04 | 0.03 |
| 10.00 | 5.01 | 1.25 | 0.63 | 0.39 | 0.28 |
| 100.00 | 49.95 | 12.47 | 6.23 | 3.89 | 2.72 |
| 300.00 | 146.98 | 85.93 | 17.64 | 10.80 | 7.41 |
| 500.00 | 236.37 | 55.87 | 26.41 | 15.60 | 10.32 |
| 700.00 | 315.59 | 71.13 | 32.05 | 18.05 | 11.39 |
| 900.00 | 384.13 | 81.95 | 34.96 | 18.64 | 11.12 |
| 10000.00 | 905.08 | 40.92 | 3.70 | 0.42 | 0.05 |

Table 1: *Expectations of size indices under* $N = 1000$ (Theorem 5)

In particular,

$$\mathrm{E}(S_i) = \frac{K_{N-i-1/2}(A) A^{1-i} (2i-3)!! N!}{K_{N-1/2}(A) i! (N-i)!}, \quad i = 1, 2, \ldots, N. \tag{15}$$

Table 1 summarizes values of $\mathrm{E}(S_i)$'s for $i = 1, 2, \ldots, 5$, with various parameter values given $N = 1000$. By the fact that $\mathrm{E}(U_N) = \sum_{i=1}^{N} \mathrm{E}(S_i)$, the following proposition holds from (15). Higher moments of $U_N$ can be obtained in an analogous way.

**Proposition 1** *Suppose that a random variable* $U_N$ *is subject to the distribution* (13)*. Then its expectation is*

$$\mathrm{E}(U_N) = \frac{N!}{K_{N-1/2}(A)} \sum_{i=1}^{N} \frac{K_{N-i-1/2}(A) A^{1-i} (2i-3)!!}{i!(N-i)!}.$$

We now state the sampling distribution of (8), which is in fact the result of replacing $N, U_N$ of the population distribution by $n, u_n$. This property is remarkably convenient in applications.

**Theorem 6** *Suppose that a population consists of* $N$ *individuals that are distributed as* (8)*. If* $n$ *individuals are drawn with simple random sampling without replacement from the population, then the sampling distribution is*

$$\mathrm{P}(s_1, \ldots, s_n) = \sqrt{\frac{\pi}{2A}} \frac{n! \exp(-A)}{A^{n-u_n} K_{n-1/2}(A)} \prod_{i=1}^{n} \{\frac{(2i-3)!!}{i!}\}^{s_i} \frac{1}{s_i!}. \tag{16}$$

Suppose that $N$ objects are partitioned into classes according to a probability distribution $p_N$. A partition structure (Kingman (1978)) is a sequence $p_1, p_2, \ldots$ of distributions wherein, assuming that an object is deleted uniformly at random from the $N$ objects, the partition of the $N-1$ remaining objects is distributed according to $p_{N-1}$. The following remark is an immediate consequence of Theorem 6.

**Remark 2** *The model* (8) *has a partition structure.*

# 4 Parameter Estimation

Now that the sampling distribution (16) is given, this section treats the parameter estimation of the distribution (8). First we construct the Maximum Likelihood (ML) estimation, and second we present an approximate moment estimator.

We denote the log likelihood of (16) by

$$L = -A - (n - u_n + \frac{1}{2}) \log A - \log K_{n-1/2}(A) + Const.$$

In the following we will use this notation:

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

It is widely known, see Seshadri (1999, p.125) for instance, that

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

The derivative of $L$:

$$\frac{\partial L}{\partial A} = -1 - (2n - u_n)\frac{1}{A} + R_{n-1/2}(A)$$

is hence easy to calculate. The ML estimator is the unique solution of $\partial L/\partial A = 0$; Remark 1 validates the uniqueness, based on the property of the exponential family discussed by Lehmann (1991, p.417) for example.

It requires a numerical method to solve the likelihood equation. We can utilize the second derivative:

$$\frac{\partial^2 L}{\partial A^2} = R^2_{n-1/2}(A) - \frac{2n}{A} R_{n-1/2}(A) + (2n - u_n)\frac{1}{A^2} - 1$$

for the Newton-Raphson method.

A moment estimator is useful also for the starting value of the Newton-Raphson procedure, but an exact one is inconvenient to compute because of the modified Bessel function. Therefore the author proposes to employ an approximate estimator. Under the distribution (9) in Theorem 2,

$$\mathrm{E}(U) = A(1 - \sqrt{1 - \theta}),$$

and

$$\frac{4\,\mathrm{E}(S_2)}{\mathrm{E}(S_1)} = \theta.$$

The solution of these equations is

$$A = \frac{\mathrm{E}(U)}{1 - \sqrt{1 - 4\mathrm{E}(S_2)/\mathrm{E}(S_1)}},$$

which leads to the following approximate moment estimator:

$$\tilde{A} = u_n/(1 - \sqrt{1 - 4s_2/s_1}\,). \tag{17}$$

In real data sets, however, $4s_2/s_1$ can be larger than unity. We can simply use $u_n$ itself as an estimate of $A$ in such a case; see an example in the next section.

| | 23 July 1946 | | | 9 August 1946 | | |
|---|---|---|---|---|---|---|
| Individ. | obs. | fit | LS | obs. | fit | LS |
| 1 | 18 | 22.79 | 15.1 | 36 | 40.52 | 29.8 |
| 2 | 8 | 5.67 | 7.0 | 8 | 9.91 | 13.1 |
| 3 | 4 | 2.82 | 4.4 | 8 | 4.84 | 7.6 |
| 4 | 3 | 1.76 | 3.3 | 6 | 2.96 | 5.1 |
| 5 | 0 | 1.23 | | 4 | 2.02 | |
| 6 | 2 | 0.91 | | 2 | 1.48 | |
| 7 | 0 | 0.72 | | 1 | 1.14 | |
| 8 | 1 | 0.58 | | 1 | 0.90 | |
| 9 | 1 | 0.48 | | 0 | 0.73 | |
| 10+ | 4 | 5.04 | | 4 | 5.50 | |

Table 2: *Frequencies of Lepidoptera at Rothamsted Experimental Station (Williams, 1964)*

## 5 Application Results and Conclusion

Before conclusion, we demonstrate the applicability of (8) as a population model, adopting classical data from Williams (1964, p.32): two catches of Macro-Lepidoptera caught on single nights in light trap "B" at Rothamsted Experimental Station. We estimate the value of the parameter $A$, and calculate each $\mathrm{E}(s_i)$ under the estimate of $A$ as a fitted value. These applications do not involve inference on a population such as the estimation of the total number of species in a population. Nevertheless, we can employ basically the same method for population inference with a slight change, i.e., the expectations of size indices are calculated given $N$ instead of $n$.

The first set is on the night of 23-24 July 1946; there were 219 moths ($n$) belonging to 42 species ($u_n$). The ML estimate $\hat{A}$ is 45.762 in this case. Because $4s_2/s_1 > 1$, an approximate estimate $\tilde{A}$ by (17) is taken to be $u = 42.0$. The second set is of 9-10 August; there were 242 moths of 70 species. The ML estimate $\hat{A} = 82.878$, with an approximate estimate $\tilde{A} = 105.0$. Table 2 shows more detail of the fit; "Individ." indicates the number of individuals ($i$); "obs." indicates the observed number of species of $i$ individuals; "fit" indicates $E(s_i)$'s given $n$ and estimated parameter value $\hat{A}$; "LS" indicates fitted values of a logarithmic series curve by Watson, for comparison.

The model (8) seems to provide reasonable fits of size indices. Needless to say, there is no simple model that can describe all kinds of data. It is thus important to investigate various models, each of which can describe different variations of populations. The proposed distribution (8) broadens the scope of population modeling. We have seen some similarities between (8) and the Ewens sampling formula; further similarity will be investigated in our subsequent works.

## Appendix

**Proof of Theorem 1** Let us rewrite the right hand side of (7) as

$$C_1 \times C_2 \times C_3,$$

where

$$C_1 = \frac{J!}{(J-U_N)!J^{U_N}},$$

$$C_2 = \frac{N!}{J^{N-U_N+1/2}K_{N-1/2}(J\alpha)} \prod_{i=1}^{N} \{\frac{K_{i-1/2}(\alpha)}{i!}\}^{S_i} \frac{1}{S_i!}$$

and

$$C_3 = (\sqrt{\frac{\pi}{2\alpha}} \exp(-\alpha))^{J-U_N} (\sqrt{\frac{2\alpha}{\pi}})^{J-1} = (\sqrt{\frac{\pi}{2\alpha}})^{1-U_N} \exp(-J\alpha + \alpha U_N).$$

Because $C_1 \to 1$, it suffices to show that $C_2 \times C_3$ converges to the limit.

As stated by Jørgensen (1982, p.171), for $\gamma > 0$

$$K_\gamma(\alpha) \to \Gamma(\gamma) 2^{\gamma-1} \alpha^{-\gamma} \tag{18}$$

as $\alpha \to 0$. Using this result, we obtain

$$\begin{aligned}
C_2 &\to \frac{N!}{J^{N-U_N+1/2}K_{N-1/2}(A)} \prod_{i=1}^{N} \{\frac{\Gamma(i-1/2)2^{i-3/2}\alpha^{1/2-i}}{i!}\}^{S_i} \frac{1}{S_i!} \\
&= 2^{N-\frac{3U_N}{2}} \alpha^{-N+\frac{U_N}{2}} \frac{N!}{J^{N-U_N+1/2}K_{N-1/2}(A)} \prod_{i=1}^{N} \{\frac{\Gamma(i-1/2)}{i!}\}^{S_i} \frac{1}{S_i!} \\
&= 2^{N-\frac{3U_N}{2}} \alpha^{-N+\frac{U_N}{2}} \frac{N!}{J^{N-U_N+1/2}K_{N-1/2}(A)} \prod_{i=1}^{N} \{\frac{2^{1-i}\sqrt{\pi}(2i-3)!!}{i!}\}^{S_i} \frac{1}{S_i!} \\
&= (\sqrt{\frac{\pi}{2}})^{U_N} \alpha^{-N+\frac{U_N}{2}} \frac{N!}{J^{N-U_N+1/2}K_{N-1/2}(A)} \prod_{i=1}^{N} \{\frac{(2i-3)!!}{i!}\}^{S_i} \frac{1}{S_i!}.
\end{aligned}$$

The limit of $C_3$ is obviously

$$(\sqrt{\frac{\pi}{2\alpha}})^{1-U_N} \exp(-A).$$

Hence we prove the theorem by multiplying these limits. Q.E.D.

**Proof of Theorem 2** The following argument is analogous to that of Hoshino and Takemura (1998), where the logarithmic series model was discussed in detail.

First we derive the probability generating function $G(z_1, z_2, \ldots)$ of (3). If $J = 1$, then

$$\begin{aligned}
G_1(z_1, z_2, \ldots) &= \mathrm{E}[\prod_{i=1}^{\infty} z_i^{S_i}] \\
&= \sum_{i=1}^{\infty} (z_i - 1)\mathrm{P}(F_1 = i) + 1 \\
&= \sum_{i=1}^{\infty} (z_i - 1)\sqrt{\frac{2\alpha}{\pi}} \exp(\alpha\sqrt{1-\alpha}) \frac{(\alpha\theta/2)^i}{i!} K_{i-1/2}(\alpha) + 1.
\end{aligned}$$

By the independence of $F_j$'s, the joint probability generating function for general $J$ is expressed as $G(z_1, z_2, \ldots) = G_1(z_1, z_2, \ldots)^J$. Now we consider the limiting process of $J \to \infty$ where $J\alpha = A$. Using (18),

$$\begin{aligned}
&[\sum_{i=1}^{\infty}(z_i - 1)\sqrt{\frac{2\alpha}{\pi}}\exp(\alpha\sqrt{1-\alpha})\frac{(\alpha\theta/2)^i}{i!}K_{i-1/2}(\alpha) + 1]^J \\
&\quad\to\quad [1 + \frac{1}{J}J\sum_{i=1}^{\infty}(z_i - 1)\sqrt{\frac{2\alpha}{\pi}}\frac{(\alpha\theta/2)^i}{i!}\Gamma(i - 1/2)2^{i-3/2}\alpha^{-i+1/2}]^J \\
&\quad\to\quad \exp(\sum_{i=1}^{\infty}(z_i - 1)A(\frac{\theta}{2})^i\frac{(2i-3)!!}{i!}) = \exp(\sum_{i=1}^{\infty}(z_i - 1)\tau_i). \qquad (19)
\end{aligned}$$

Equation (19) coincides with the joint probability generating function of independent Poisson variables $S_i$, $i = 1, 2, \ldots$, with mean $\mathrm{E}(S_i) = \tau_i$.

We can obtain the distribution of the population size $N$ under (9) through the limiting argument over (6) of the IGP population model (3). Because equation (6) remains unchanged when $J\alpha$ is fixed, (6) is the population size distribution of (9). The conditional model given $N$ is thus the result of dividing (9) by (6), which equals (8) using (1) and (2). Conversely (8) multiplied by (6) is (9). Q.E.D.

**Proof of Theorem 3** We can obtain (12) as a result of applying (18) to (11):

$$\frac{1}{2\sqrt{\pi}}\frac{\alpha\exp(\alpha)}{\exp(\alpha(1-\sqrt{1-\theta})) - 1}\frac{\theta^x\Gamma(x - 1/2)}{x!},$$

where

$$\frac{\alpha\exp(\alpha)}{\exp(\alpha(1-\sqrt{1-\theta})) - 1} = \frac{\alpha(1 + \alpha + O(\alpha^2))}{\alpha(1-\sqrt{1-\theta}) + O(\alpha^2)} \to \frac{1}{1-\sqrt{1-\theta}}$$

as $\alpha \to 0$. Since $\Gamma(x - 1/2)/(2\sqrt{\pi}) = (2x - 3)!!/2^x$, the theorem holds. Q.E.D.

**Proof of Theorem 4** From the definition (8),

$$K_{N-1/2}(A)\mathrm{P}(U_N = v) = \sum_{U_N=v}\sqrt{\frac{\pi}{2A}}\frac{N!\exp(-A)}{A^{N-U_N}}\prod_{i=1}^{N}\{\frac{(2i-3)!!}{i!}\}^{S_i}\frac{1}{S_i!}.$$

Then using (4), we obtain

$$(\sum_{i=0}^{N-1}\frac{(N-1+i)!}{(N-1-i)!i!}(2A)^{-i})\mathrm{P}(U_N = v) = \sum_{U_N=v}\frac{N!}{A^{N-U_N}}\prod_{i=1}^{N}\{\frac{(2i-3)!!}{i!}\}^{S_i}\frac{1}{S_i!}.$$

Because $\sum_{v=1}^{N}\mathrm{P}(U_N = v) = 1$,

$$\sum_{i=0}^{N-1}\frac{(N-1+i)!}{(N-1-i)!i!}(2A)^{-i} = \sum_{v=1}^{N}\sum_{U_N=v}\frac{N!}{A^{N-U_N}}\prod_{i=1}^{N}\{\frac{(2i-3)!!}{i!}\}^{S_i}\frac{1}{S_i!}.$$

By comparing the coefficient of $A$ between the left hand side and the right hand side of the equation above, we have

$$\sum_{U_N=v} N!\{\frac{(2i-3)!!}{i!}\}^{S_i}\frac{1}{S_i!} = \frac{(N-1+N-v)!}{(N-1-N+v)!(N-v)!}\,2^{v-N}.$$

It then leads to (13). Q.E.D.

**Proof of Theorem 5** Let us write

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

We can show (14) by the fact that $\sum_{\boldsymbol{S}\in\mathcal{S}(N)} \mathrm{P}(\boldsymbol{S}) = 1$:

$$\begin{aligned}
\mathrm{E}(\prod_{i=1}^{N} S_i^{(r_i)}) &= \sum_{\boldsymbol{S}\in\mathcal{S}(N)} \sqrt{\frac{\pi}{2A}}\frac{N!\exp(-A)}{A^{N-U_N}K_{N-1/2}(A)}\prod_{i=1}^{N}\{\frac{(2i-3)!!}{i!}\}^{S_i-r_i}\frac{1}{(S_i-r_i)!}\{\frac{(2i-3)!!}{i!}\}^{r_i} \\
&= \frac{N!A^{r-R}K_{N-R-1/2}(A)}{(N-R)!K_{N-1/2}(A)} \sum_{\boldsymbol{S}\in\mathcal{S}(N)} \sqrt{\frac{\pi}{2A}}\frac{(N-R)!\exp(-A)}{A^{N-R-(U_N-r)}K_{N-R-1/2}(A)} \times \\
&\quad \prod_{i=1}^{N}\{\frac{(2i-3)!!}{i!}\}^{S_i-r_i}\frac{1}{(S_i-r_i)!}\{\frac{(2i-3)!!}{i!}\}^{r_i} \\
&= \frac{N!A^{r-R}K_{N-R-1/2}(A)}{(N-R)!K_{N-1/2}(A)}\prod_{i=1}^{N}\{\frac{(2i-3)!!}{i!}\}^{r_i} \times \\
&\quad \sum_{\boldsymbol{S}\in\mathcal{S}(N-R)} \sqrt{\frac{\pi}{2A}}\frac{(N-R)!\exp(-A)}{A^{N-R-(U_{N-R})}K_{N-R-1/2}(A)}\prod_{i=1}^{N}\{\frac{(2i-3)!!}{i!}\}^{S_i}\frac{1}{S_i!} \\
&= \frac{N!A^{r-R}K_{N-R-1/2}(A)}{(N-R)!K_{N-1/2}(A)}\prod_{i=1}^{N}\{\frac{(2i-3)!!}{i!}\}^{r_i}.
\end{aligned}$$

Q.E.D.

**Proof of Theorem 6** This result is a direct consequence of Lemma 1 in Takemura (1999). It assures that, if the prior distribution of the values of $N$ individuals is exchangeable with respect to the individuals, the marginal distribution of the values of $n$ individuals drawn with simple random sampling without replacement coincides with the prior distribution of values of $n$ individuals directly drawn from the superpopulation. Since our model (8) does not depend on labels of individuals, the theorem holds. Q.E.D.

## References


[1] Anscombe, F.J. (1950). Sampling theory of the negative binomial and logarithmic series distributions. *Biometrika*, **37**, 358–382.

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

[3] Bunge, J. and Fitzpatrick, M. (1993). Estimating the number of species: a review. *Journal of the American Statistical Association*, **88**, 364–373.

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

[5] 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.

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

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

[8] 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.

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

[10] Hoshino, N. (2001b). On a conditional inverse Gaussian-Poisson distribution. *Technical Report CMU-CALD-01-102*, School of Computer Science, Carnegie Mellon University.

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

[12] Johnson, N.L., Kotz, S. and Kemp, A.W. (1993). *Univariate Discrete Distributions*, 2nd, Wiley, New York.

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

[14] Kingman, J.F. (1978). Random partitions in population genetics. *Proceedings of the Royal Society of London*, A, **361**, 1–20.

[15] Lehmann, E.L. (1991). *Theory of point estimation*. Wadsworth, California.

[16] Noack, A. (1950). A class of random variables with discrete distributions. *Annals of Mathematical Statistics*, **21**, 127–132.

[17] Sampford, M.R. (1955). The truncated negative binomial distribution. *Biometrika*, **42**, 58–69.

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

[19] Sibuya, M. (1991). A cluster-number distribution and its application to the analysis of homonyms. *Japanese Journal of Applied Statistics*, **20**, 139–153 (in Japanese).

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

[21] Sichel, H.S. (1975). On a distribution law for word frequencies. *Journal of the American Statistical Association*, **70**, 542–547.

[22] 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.

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

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

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

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