Engen's Extended Negative Binomial Model Revisited

Nobuaki Hoshino*
Faculty of Economics, Kanazawa University

August 7, 2004

Abstract

The present article shows that a limiting argument that is essentially the law of small numbers produces a proper discrete multivariate distribution from any generalized Poisson distribution. Based on this result, Engen's Extended Negative Binomial (ENB) model is derived from the Poisson-Pascal distribution, which is a generalization of the inverse Gaussian-Poisson distribution. The ENB model is also derived from Sichel's generalized inverse Gaussian-Poisson distribution. The application of the ENB model is discussed thereto.

Keywords: Compound Poisson, Conditional inverse Gaussian Poisson, Infinitely divisible, Random clustering, Species abundance

1 Introduction

Engen (1974) proposed the Extended Negative Binomial (ENB) model to describe the population structure of frequencies of species. A population model of this type is called a stochastic abundance model (Engen (1978)) and is statistically a distribution over nonnegative integers. This kind of population modeling has extensive applicability and is thus an important subject. Many linguists, for instance, have applied population models to word frequencies, and recently statistical disclosure control demands continuous development in this modeling; see Hoshino (2001) for a brief survey.

The ENB model was, however, not clearly specified enough to attract many statisticians' interest, as explained in Section 1.1. Consequently, there remain many points to be clarified. In order to elucidate relationships among the ENB model and other models, the present article in Section 2 shows a limiting property of a population model that consists of generalized Poisson distributions in the sense of Johnson et al. (1993, p.351). An instance of this distribution called Poisson-Pascal will result in the ENB model. It is also shown that the same type of limiting produces the ENB model from Sichel (1971)'s generalized inverse Gaussian-Poisson distributions. The present article demonstrates the applicability of the ENB model in Section 3. Concluding remarks are given in Section 4.

Because the class of generalized Poisson distributions contains most of distributions that have been used to describe frequency data, these discussions bring profound understanding on models for count data.

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

1.1 Background

In the following, $\mathbb{N}$ denotes the set of natural numbers; $\mathbb{N}_0$ denotes the set of nonnegative integers. For an arbitrary nonnegative integer J, the set of positive integers from one to J, or $\{1, 2, \ldots, J\}$, is denoted by $\mathbb{N}(J)$. Consider a population consisting of J cells (groups, species, words), each of which is uniquely indexed by j, where $j \in \mathbb{N}(J)$. The j-th cell contains F_j individuals, where $F_j \in \mathbb{N}_0$, and the total number of individuals is denoted by $N = \sum_{j=1}^{J} F_j$.

Let S_i denote the number of cells of size i. More specifically,

$$S_i = \sum_{j=1}^{J} I(F_j = i), \quad i \in \mathbb{N}_0,$$

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, $(S_0, S_1, \ldots)$ are called size indices (Sibuya (1993)) or frequencies of frequencies (Good (1953)).

Obviously the S_i are nonnegative integers that satisfy

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

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

It is noteworthy that J is the total number of cells including empty cells, which may correspond to unseen or extinct species. In the following

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

denotes the number of non-empty cells.

A typical assumption of a model for count data is that $F_j, j \in \mathbb{N}(J)$, is independently and identically distributed over nonnegative integers. Then, as explained in Appendix A, size indices are multinomially distributed over ${\mathbb{N}_0}^{\infty}$:

$$\mathrm{P}(S_1 = t_1, S_2 = t_2, \ldots) = J! \prod_{i=0}^{\infty} \frac{\mathrm{P}(F_1 = i)^{t_i}}{t_i!}, \quad t_0 = J - \sum_{i=1}^{\infty} t_i \geq 0. \tag{3}$$

For example, suppose that $F_j, j \in \mathbb{N}(J)$, is independently identically distributed as the negative binomial distribution:

$$\mathrm{P}(F_j = y) = \frac{(1-\theta)^{\gamma} \theta^{y} \Gamma(y+\gamma)}{\Gamma(\gamma) y!}, \quad y \in \mathbb{N}_0,\ 0 < \theta < 1, 0 < \gamma. \tag{4}$$

The joint distribution of size indices results in

$$\mathrm{P}(S_1 = t_1, S_2 = t_2, \ldots) = J!(\prod_{i=0}^{\infty}(\frac{\Gamma(i+\gamma)}{\Gamma(\gamma)i!}(1-\theta)^{\gamma}\theta^{i})^{t_i}\frac{1}{t_i!}, \quad t_0 = J - \sum_{i=1}^{\infty} t_i \geq 0. \tag{5}$$

The expectation of a size index is

$$\mathrm{E}(S_i) = J\frac{(1-\theta)^{\gamma}\theta^{i}\Gamma(i+\gamma)}{\Gamma(\gamma)\, i!}, \quad i \in \mathbb{N}_0, \tag{6}$$

under (5). The population size N becomes a random variable, and let us restrict its expectation to the actual population size N_0, which is usually given in practice. Under (5), the restriction $\mathrm{E}(N) = \sum_{i=1}^{\infty} i\mathrm{E}(S_i) = N_0$ is equivalent to

$$J\gamma = \frac{N_0(1-\theta)}{\theta}. \tag{7}$$

An actual population often consists of very large number of cells. It is thus reasonable to consider the limit of a model as

$$J \to \infty, \text{ where } \mathrm{E}(N) = N_0 \text{ fixed}. \tag{8}$$

Anscombe (1950) pointed out that applying (8) to (5) produces the logarithmic series model:

$$\mathrm{P}(S_1 = t_1, S_2 = t_2, \ldots) = \prod_{i=1}^{\infty}\frac{{\lambda_i}^{t_i}\exp(-\lambda_i)}{t_i!}, \tag{9}$$

where

$$\lambda_i = \frac{N_0(1-\theta)}{\theta}\frac{\theta^{i}}{i}.$$

Under (9), each S_i is independently subject to the Poisson distribution with mean λ_i, which is henceforth denoted by $Po(\lambda_i)$. The model (9) is named after the logarithmic series distribution (Fisher et al. (1943)), since the series of λ_i is based on the same series expansion.

Using the restriction (7), we can rewrite (6) as

$$\mathrm{E}(S_i) = \frac{N_0}{\theta}\frac{(1-\theta)^{\gamma+1}\theta^{i}\Gamma(i+\gamma)}{\Gamma(\gamma+1)i!} := \tau(i;\gamma,\theta). \tag{10}$$

Observing (10), Engen (1974) claimed for $i \in \mathbb{N}$ that the natural lower bound of γ is -1 in contrast to the logarithmic series model where $\gamma \to 0$; the "extended" part of the ENB model is this newly introduced area of $-1 < \gamma < 0$. However, only the expectation of a size index (10) was given, and the joint distribution of size indices was not specified. See Section 3.4 of Engen (1978) or Section 5.12.2 of Johnson et al. (1993) for more information. To avoid confusion, the ENB model is discriminated from the Extended (truncated) Negative Binomial distribution:

$$\mathrm{P}(x) = \frac{-\gamma}{1-(1-\theta)^{-\gamma}}\frac{\theta^{x}\Gamma(x+\gamma)}{\Gamma(\gamma+1)x!} \propto \tau(x;\gamma,\theta) \quad , x \in \mathbb{N},\, 0 < \theta < 1, -1 < \gamma < 0. \tag{11}$$

When γ is positive, (11) is the usual truncated negative binomial distribution. Sichel (1997) remarked that the ENB distribution fits a good number of observed species frequencies rather

well as its skewness lies somewhere between that of the logarithmic series distribution and the lognormal-Poisson distribution. This fact accords with Engen's claim that the ENB model can describe various actual populations.

Recently, Hoshino (2002) obtained a population model:

$$\mathrm{P}(S_1 = t_1, S_2 = t_2, \ldots) = \prod_{i=1}^{\infty} \frac{\exp(-\tau(i; -1/2, \theta))\tau(i; -1/2, \theta)^{t_i}}{t_i!}, \tag{12}$$

where

$$\tau(i; -1/2, \theta) = \frac{2N_0\sqrt{1-\theta}}{\theta}(\frac{\theta}{2})^i \frac{(2i-3)!!}{i!} = \frac{N_0}{\theta}\frac{\sqrt{1-\theta}\theta^i}{\Gamma(1/2)}\frac{\Gamma(i-1/2)}{\Gamma(i+1)},$$

$(-1)!! = 1$ and $(2i-3)!! = (2i-3)(2i-5)\cdots 1$. In (12), each S_i is independently distributed as $Po(\tau(i; -1/2, \theta))$. Hence (12) satisfies the restriction (10) with $\gamma = -1/2$ and can be regarded as a special case of the ENB model. Let us consider the following ENB model:

$$\mathrm{P}(S_1 = t_1, S_2 = t_2, \ldots) = \prod_{i=1}^{\infty} \frac{\exp(-\tau(i; \gamma, \theta))\tau(i; \gamma, \theta)^{t_i}}{t_i!}, \tag{13}$$

where $-1 < \gamma < 0$. Henceforth (13) is referred to as ENB(γ), and (12) is ENB($-1/2$).

The derivation of ENB($-1/2$) is as follows. Suppose that $F_j, j \in \mathbb{N}(J)$, is independently and identically distributed as the Inverse Gaussian-Poisson (IGP) distribution:

$$\mathrm{P}(F = y) = \sqrt{\frac{2\alpha}{\pi}}\exp(\alpha\sqrt{1-\theta})\frac{(\alpha\theta/2)^y}{y!}K_{y-1/2}(\alpha), \quad y \in \mathbb{N}_0,\ 0 < \alpha,\ 0 < \theta < 1, \tag{14}$$

where $K_\gamma(\cdot)$ is the modified Bessel function of the third kind of order γ; see Chap. 7.1 of Seshadri (1999) for the IGP distribution. Equation (12) is the result of applying the limiting argument (8) to this IGP population model. The next section investigates what kinds of generalizations of the IGP distribution lead to ENB(γ) as (8).

2 The derivation of the ENB model

This section explicates two methods each of which produces ENB(γ). The first one uses the general property of an infinitely divisible distribution over nonnegative integers. In this way, we also generalize known results about conditioning a population model on N. The second one links Sichel's generalization of the IGP distribution with the ENB model. All the proofs of theorems in this section are provided in Appendix B.

A good place to start is to examine the condition under which the limiting distribution of a size index is an independent Poisson distribution. We have seen that, if $F_j, j \in \mathbb{N}(J)$, is independently and identically distributed, size indices are subject to the multinomial distribution (3). Because the marginal distribution of the multinomial distribution is the binomial distribution, the law of small numbers applies except for S_0 (or any margin). Engen (1977) summarized this result as Lemma 1 below. The derivation of independent Poisson distributions from a finite-dimensional multinomial distribution appears, say, in Johnson et al. (1997, p.124); the set of finite number of independent Poisson distributions is called a multiple Poisson distribution there.

Lemma 1 *Let $F_j, j \in \mathbb{N}(J)$, be independently and identically distributed over $\mathbb{N}_0$. If, for each positive i, the expectation of S_i converges to a positive constant as $J \to \infty$, or*

$$\lim_{J\to\infty} J\mathrm{P}(F_j = i) = c_i, \quad i \in \mathbb{N}, \tag{15}$$

where $c_i > 0$, the limiting distribution of S_i as $J \to \infty$ is independently $Po(c_i)$.

An infinite series of size indices that are subject to independent Poisson distributions is called composed Poisson distributions; see Johnson et al. (1997, p.188). When S_i is subject to $Po(c_i)$, the probability generating function (pgf) of the number of individuals from cells of size i equals

$$\sum_{x=0}^{\infty} \frac{z^{x\cdot i} {c_i}^x \exp(-c_i)}{x!} = \exp(c_i(z^i - 1)).$$

Therefore, the pgf of N under composed Poisson distributions is expressed as

$$G_c(z) = \prod_{i=1}^{\infty} \exp(c_i(z^i - 1)).$$

If $\sum_{i=1}^{\infty} c_i = C < \infty$, we can rewrite $G_c(z)$ as

$$G_c(z) = \exp(C(g(z) - 1)),$$

where

$$g(z) = \sum_{i=1}^{\infty} \frac{c_i}{C} z^i.$$

Let c_i/C be denoted by q_i. Because q_i is positive and $\sum_{i=1}^{\infty} q_i = 1$, we can regard $g(z)$ as a pgf. The distribution defined by this $G_c(z)$ is called a generalized Poisson distribution, where $g(z)$ defines its generalizing distribution. Obviously,

$$\mathrm{E}(N) = \sum_{i=1}^{\infty} i\mathrm{E}(S_i) = C\sum_{i=1}^{\infty} iq_i := N_0.$$

The next question is to determine the distribution of F_j that satisfies the condition (15) of Lemma 1, given $\{c_i | i \in \mathbb{N}\}$; the ENB model arises when $c_i = \tau(i; \gamma, \theta)$. As a matter of course, such a distribution can not be unique. Later two distributions are shown to share the same limiting distribution, for example.

In order to determine the distribution of F_j uniquely, one may restrict the model such that the distribution of N does not change for all J, by which the distribution of N remains unchanged after the limiting argument (8) and then $\mathrm{E}(N)$ is restricted to N_0. This is possible by letting the pgf of F_j be

$$G_F(z) = \exp(\frac{C}{J}(g(z) - 1)),$$

which is again a generalized Poisson distribution. In this case, the pgf of N can be written as

$$G_c(z) = G_F(z)^J \tag{16}$$

for all J, with the result that the distribution of N is infinitely divisible. Conversely, any infinitely divisible distribution over nonnegative integers is a generalized Poisson distribution by Lévy's Theorem; see Section 12.3 of Feller (1957). It implies the following fact.

Remark 1 *Suppose that $F_j, j \in \mathbb{N}(J)$, are independently and identically subject to a proper distribution over nonnegative integers. Then the distribution of N remains unchanged for all positive J, only if F_j's are subject to a generalized Poisson distribution.*

It is thus important to elucidate the property of a model that consists of independent and identical generalized Poisson distributions. In fact, to any model of this type, the limiting argument (8) can apply.

Theorem 1 *Suppose that each $F_j, j \in \mathbb{N}(J)$, is independently and identically subject to the distribution that has the pgf:*

$$G(z) = \exp(a(g(z) - 1))), \quad 0 < a < \infty, \tag{17}$$

where

$$g(z) = \sum_{i=1}^{\infty} q_i z^i$$

is the pgf of a proper distribution over positive integers. Let $Ja = \mu$ be fixed. The limiting distribution of $S_i, i \in \mathbb{N}$, as $J \to \infty$ $(a \to 0)$ is independently $Po(q_i\mu)$.

The negative binomial distribution (4) is infinitely divisible with $a = -\gamma \log(1-\theta)$ and $g(z) = \log(1-\theta z)/\log(1-\theta)$, which defines the logarithmic series distribution:

$$q_i = -\frac{1}{\log(1-\theta)} \frac{\theta^i}{i}.$$

Thus, by letting μ equal the right hand side of (7) times $-\log(1-\theta)$, we obtain the logarithmic series model (9) as $\gamma \to 0$. Willmot (1986) noted that the IGP distribution is also infinitely divisible; let $a = \alpha(1-\sqrt{1-\theta})$ and $g(z) = (1-\sqrt{1-z\theta})/(1-\sqrt{1-\theta})$, which is the pgf of the truncated ENB distribution with $\gamma = -1/2$:

$$q_i = \frac{1}{1-\sqrt{1-\theta}} \frac{\theta^i (2i-3)!!}{2^i i!}.$$

Then, if $\mu = 2N_0(1-\sqrt{1-\theta})\sqrt{1-\theta}/\theta$, the limiting distribution is (12) as $\alpha \to 0$. See Section 8.3 of Johnson et al. (1993) for other infinite divisible distributions.

To prove Theorem 1, the author referred to Kemp (1978), where it is shown that $g(z)$ of $G_c(z)$ defines the limiting distribution of the truncated distribution of $G_c(z)$ as $C \to 0$. For instance, the logarithmic series distribution is the limit of the truncated negative binomial distribution. The limiting distribution of the truncated IGP distribution as $\alpha \to 0$ is the ENB distribution (11) with $\gamma = -1/2$.

Next we consider conditioning the population model that consists of independent generalized Poisson distributions on N. Sibuya et al. (1964) pointed out that the conditional distribution of the negative binomial model (5) given N is the Dirichlet-multinomial mixture or the negative multivariate hypergeometric distribution proposed by Mosimann (1962). Hoshino (2003) discussed the property of the conditional IGP population model given N (CIGP distribution). According to Watterson (1974), the conditional distribution of the logarithmic series model on N is the Ewens distribution (Ewens (1972)); see Hoshino and Takemura (1998) also. The conditional distribution of ENB$(-1/2)$ or the limiting CIGP distribution was derived in Hoshino

(2002), where these relationships were illustrated. This type of conditioning is of importance because fixed N is more realistic than to fix $\mathrm{E}(N)$ in application fields where a sampling frame is definite.

As for models in a broad class, the conditioning has another advantage. Let us write

$$g(z) = \sum_{i=1}^{\infty} \frac{a_i h(\theta)^i}{\eta(\theta)} z^i, \tag{18}$$

where

$$\eta(\theta) = \sum_{i=1}^{\infty} a_i h(\theta)^i, \quad a_i \geq 0,\ h(\theta) > 0.$$

The distribution defined by the pgf (18) is called a Modified Power Series (MPS) distribution (Gupta (1974)). If $h(\theta) = \theta$, which is the case of the negative binomial distribution and the IGP distribution, (18) reduces to that of a power series distribution (Noack (1955)); see Johnson et al. (1993, p.70). The following theorem states that the power parameter θ does not affect the conditional model of the Poisson distribution generalized by an MPS distribution given its total frequency. In other words, N is a sufficient statistic for θ; see Johnson et al. (1993, p.73). After conditioning on N, parameter estimation should become easier.

Theorem 2 *Suppose that each $F_j, j \in \mathbb{N}(J)$, is independently and identically subject to the distribution that has the pgf:*

$$G(z) = \exp(\alpha\eta(\theta)(g(z) - 1))), \quad 0 < \alpha < \infty,$$

where $\eta(\theta)$ and $g(z)$ are defined by (18)*. Then $G(z)$ also defines the MPS distribution:*

$$\mathrm{P}(F_j = i) = \frac{b_i h(\theta)^i}{\exp(\alpha\eta(\theta))}, \tag{19}$$

where $b_0 = 1$ and $b_{i+1} = \alpha(i+1)^{-1} \sum_{j=0}^{i} (i+1-j) a_{i+1-j} b_j$.

The conditional model given $N = \sum_{j=1}^{J} F_j$ is expressed as

$$\mathrm{P}(F_1 = g_1, F_2 = g_2, \ldots, F_J = g_J | N = n) = \prod_{j=1}^{J} b_{g_j} / d_n,$$

or

$$\mathrm{P}(S_1 = t_1, S_2 = t_2, \ldots, S_n = t_n | N = n) = \frac{J!}{(J-v)! d_n} \prod_{i=1}^{n} \frac{{b_i}^{t_i}}{t_i!}, \quad v = \sum_{i=1}^{n} t_i, \tag{20}$$

where $d_0 = 1$ and $d_{i+1} = J\alpha(i+1)^{-1} \sum_{j=0}^{i} (i+1-j) a_{i+1-j} d_j$.

When $J\alpha$ is fixed at μ, the limiting distribution of (20) *as $J \to \infty$ is*

$$\mathrm{P}(S_1 = t_1, S_2 = t_2, \ldots, S_n = t_n | N = n) = \frac{\mu^v}{d_n} \prod_{i=1}^{n} \frac{{a_i}^{t_i}}{t_i!}.$$

Now, ENB(γ) is derived based on Theorem 1. The ENB distribution (11) has the following pgf:

$$g(z) = \frac{1-(1-z\theta)^{-\gamma}}{1-(1-\theta)^{-\gamma}},$$

by which we obtain a generalized Poisson distribution defined by this pgf:

$$G(z) = \exp(\alpha\{(1-\theta)^{-\gamma} - (1-z\theta)^{-\gamma}\}), \tag{21}$$

where $0 < \alpha, 0 < \theta < 1$ and $-1 < \gamma < 0$. Actually, γ can be positive, and (21) reduces to that of the Poisson distribution when $\gamma = -1$. However, the present article only considers the aforementioned parameter space. Let the distribution of $F_j, j \in \mathbb{N}(J)$, be defined by (21). Then

$$\mathrm{E}(F_j) = -\alpha\gamma\,\theta(1-\theta)^{-\gamma-1}, \quad j \in \mathbb{N}(J),$$

and the relationship that $\mathrm{E}(N) = N_0$ is equivalent to the restriction:

$$J\alpha = -\frac{N_0(1-\theta)^{1+\gamma}}{\gamma\theta}. \tag{22}$$

When $a = \alpha(1-(1-\theta)^{-\gamma})$ and μ equals the right hand side of (22), Theorem 1 produces ENB(γ). In summary, the following result holds.

Proposition 1 *Suppose that $F_j, j \in \mathbb{N}(J)$, is independently subject to the identical distribution that has the pgf* (21). *Let* $\mathrm{E}(N)$ *be fixed at N_0. The limiting distribution of $(S_1, S_2, \ldots)$ as $J \to \infty\,(\alpha \to 0)$ is then* ENB(γ).

The distribution (21) is called Poisson-Pascal and reviewed by Johnson et al. (1993, p.382), in which the case of positive γ is solely considered, though. It was Willmot (1989) who pointed out that negative γ larger than -1 is valid. If we allow the first moment of the Poisson-Pascal distribution to be infinite, θ can be unity, where (21) reduces to that of the discrete stable distribution (Steutel and van Harn (1979)) and the ENB distribution (11) reduces to the Sibuya distribution (Sibuya (1979)). It is observable that the ENB distribution is a power-series-distributionized Sibuya distribution and the Poisson-Pascal distribution is a power-series-distributionized discrete stable distribution.

The probability function of the Poisson-Pascal distribution is generally complicated. However, because the ENB distribution (11) belongs to the class of MPS distributions, Theorem 2 assures us of the following result.

Proposition 2 *Suppose that $F_j, j \in \mathbb{N}(J)$, is independently subject to the identical distribution that has the pgf* (21). *Then*

$$\mathrm{P}(F_j = i) = \frac{D(i;\alpha,\gamma)\theta^i}{\exp(\alpha(1-(1-\theta)^{-\gamma}))}, \quad i \in \mathbb{N}_0, \tag{23}$$

where $D(0;\alpha,\gamma) = 1$ and

$$D(i+1;\alpha,\gamma) = \frac{\alpha}{i+1}\sum_{j=0}^{i}\frac{-\gamma\Gamma(i+1-j+\gamma)}{\Gamma(\gamma+1)(i-j)!}D(j;\alpha,\gamma). \tag{24}$$

The conditional model given $N = \sum_{j=1}^{J} F_j$ *is expressed as*

$$\mathrm{P}(F_1 = g_1, F_2 = g_2, \ldots, F_J = g_J | N = n) = \frac{1}{D(n; J\alpha, \gamma)} \prod_{j=1}^{J} D(g_j; \alpha, \gamma)$$

or

$$\mathrm{P}(S_1 = t_1, S_2 = t_2, \ldots, S_n = t_n | N = n) = \frac{J!}{D(n; J\alpha, \gamma)(J-v)!} \prod_{i=1}^{n} \frac{D(i; \alpha, \gamma)^{t_i}}{t_i!}, \tag{25}$$

where $v = \sum_{i=1}^{n} t_i$.

The limiting distribution of (25) *as* $J \to \infty$ *when* $J\alpha = \mu$ *is*

$$\mathrm{P}(S_1 = t_1, S_2 = t_2, \ldots, S_n = t_n | N = n) = \left\{ \frac{-\gamma\mu}{\Gamma(\gamma+1)} \right\}^{v} \frac{1}{D(n; \mu, \gamma)} \prod_{i=1}^{n} \left\{ \frac{\Gamma(i+\gamma)}{i!} \right\}^{t_i} \frac{1}{t_i!}. \tag{26}$$

When μ *equals the right hand side of* (22)*, equation* (26) *is the conditional distribution of ENB*(γ) *given* N.

We may recall that (21) reduces to that of the IGP distribution when $\gamma = -1/2$. Hence (25) reduces to the CIGP distribution when $\gamma = -1/2$, and (26) reduces to the limiting CIGP distribution. When $\gamma \to 0$, (26) corresponds to the Ewens distribution. These special cases are very simple.

Remark 2 *The distribution* (26) *belongs to an exponential family, when* γ *is fixed and* μ *is seen as the unique parameter. Then* U *is its sufficient statistic, as the Ewens distribution* ($\gamma = 0$) *and the limiting CIGP distribution* ($\gamma = -1/2$).

Let us investigate the number $D(i; \alpha, \gamma)$ defined by the recursion (24). Its generating function appears to be

$$f(z; \alpha, \gamma) = \exp(\alpha(1 - (1-z)^{-\gamma}))$$

because of (23). Charalambides and Singh (1988, eq. 3.19) evaluated it as

$$f(z; \alpha, \gamma) = \sum_{i=0}^{\infty} \sum_{j=1}^{i} C(i, j, -\gamma)(-\alpha)^j (-z)^i \frac{1}{i!}, \tag{27}$$

where $C(i, j, -\gamma)$ is the C-number, which is a generalized Stirling number; see Charalambides and Singh (1988) for its detailed review. In the domain of $-1 < \gamma < 0$, the expression of the C-number is not simple except for the case of $\gamma = -1/2$. Because equation (27) implies that

$$D(i; \alpha, \gamma) = \sum_{j=1}^{i} \alpha^j (-1)^{i+j} C(i, j, -\gamma) \frac{1}{i!},$$

it is unlikely that $D(i; \alpha, \gamma)$ can be generally expressed in a simple form.

Professor H. Yamato suggested the following evaluation of the distribution of U under (26). See Hoshino (2002) for its simple case of $\gamma = -1/2$. Let us rewrite the right hand side of (26) as

$$\frac{(-\gamma\mu)^v}{D(n; \mu, \gamma)} \prod_{i=1}^{n} \left\{ \frac{(1+\gamma)^{[i-1]}}{i!} \right\}^{t_i} \frac{1}{t_i!} = \frac{\mu^v}{D(n; \mu, \gamma)} (-1)^{n-v} \prod_{i=1}^{n} \binom{-\gamma}{i}^{t_i} \frac{1}{t_i!}, \tag{28}$$

where $(1+\gamma)^{[i-1]} = \Gamma(i+\gamma)/\Gamma(1+\gamma)$. Charalambides and Singh (1988, eq. 3.24) showed

$$C(n, v, -\gamma) = \sum_{t_1+t_2+\cdots+t_n=v} n! \prod_{i=1}^{n} \binom{-\gamma}{i}^{t_i} \frac{1}{t_i!}, \tag{29}$$

where the summation is taken over all partitions of n into v parts under $\sum it_i = n$. Consequently,

$$\mathrm{P}(U = v|N = n) = \frac{\mu^v}{D(n; \mu, \gamma)} (-1)^{n-v} \frac{1}{n!} C(n, v, -\gamma),$$

assuming (26).

The above derivation follows Yamato et al. (2001), who evaluated the distribution of U under a generalized Ewens distribution called Pitman's sampling formula. Pitman (1995) defined it as

$$\mathrm{P}(S_1 = t_1, \ldots, S_n = t_n|N = n) = n! \frac{\theta^{(v;\gamma)}}{\theta^{[n]}} \prod_{i=1}^{n} \left\{ \frac{(1+\gamma)^{[i-1]}}{i!} \right\}^{t_i} \frac{1}{t_i!},$$

where $v = \sum_{i=1}^{n} t_i$ and $\theta^{(v;\gamma)} = \theta(\theta-\gamma)\cdots(\theta-(v-1)\gamma)$. The parameter space of $-1 < \gamma < 0$ is valid for $\theta > \gamma$, and U is sufficient for θ because

$$\mathrm{P}(S_1 = t_1, S_2 = t_2, \ldots, S_n = t_n|U = v, N = n) = \frac{n!}{C(n, v, -\gamma)} \prod_{i=1}^{n} \binom{-\gamma}{i}^{t_i} \frac{1}{t_i!} \tag{30}$$

under Pitman's sampling formula. The following fact has been obvious.

Remark 3 *The conditional distribution of* (26) *given* U *and* N *is the same as* (30) *of Pitman's sampling formula.*

The restriction (16) was expedient to determine the distribution of F_j uniquely. Removing this restriction allows us to show that another model converges in distribution to ENB(γ) by the limiting argument (8). Sichel (1971) proposed the Generalized IGP (GIGP) distribution or Sichel distribution:

$$\mathrm{P}(F = y) = \frac{(1-\theta)^{\gamma/2}(\alpha\theta/2)^y}{K_\gamma(\alpha\sqrt{1-\theta})y!} K_{y+\gamma}(\alpha), \quad y \in \mathbb{N}_0, 0 < \theta < 1, 0 < \alpha, \tag{31}$$

which equals the IGP distribution (14) when $\gamma = -1/2$. The limiting distribution of (31) as $\alpha \to 0$ is the negative binomial distribution when γ is positive. Sichel (1992) showed for $-1 < \gamma < 0$ that the limiting distribution of the truncated GIGP distribution as $\alpha \to 0$ is the ENB distribution. As regards the GIGP model:

$$\mathrm{P}(F_1 = g_1, F_2 = g_2, \ldots, F_J = g_J) = \prod_{j=1}^{J} \frac{(1-\theta)^{\gamma/2}(\alpha\theta/2)^{g_j}}{K_\gamma(\alpha\sqrt{1-\theta})g_j!} K_{g_j+\gamma}(\alpha), \tag{32}$$

the limiting distribution as $\alpha \to 0\,(\gamma > 0)$ is the negative binomial model, which becomes the logarithmic series model by the limiting argument (8). We are interested in the case of $-1 < \gamma < 0$ on (32), which becomes ENB(γ).

Theorem 3 *If* $-1 < \gamma < 0$, *under the restriction that* $\mathrm{E}(N) = N_0$, *the limiting distribution of size indices of* (32) *is* ENB(γ) *as* $J \to \infty$ *and* $\alpha \to 0$.

The relationships shown in this section are summarized in Figure 1.

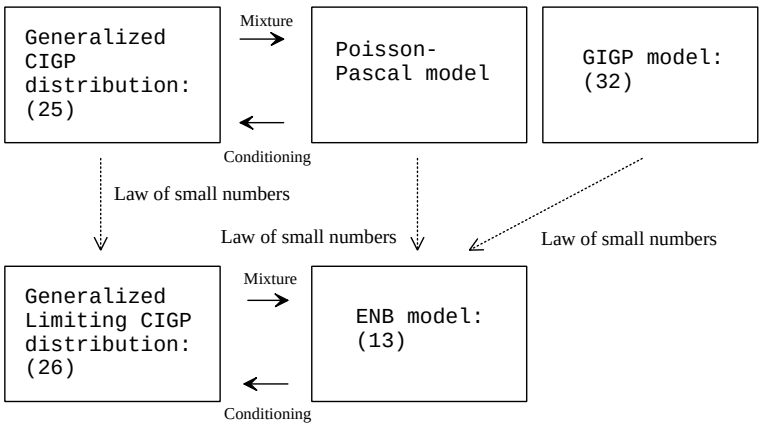

Figure 1: Models relating to ENB(γ)

3 Applying the ENB model

3.1 Parameter estimation

This section discusses the Maximum Likelihood (ML) estimation of ENB(γ). For $i = 1, 2, \ldots,$ an observed size index S_i is denoted by t_i, and $v = \sum_{i=1}^{\infty} t_i, n = \sum_{i=1}^{\infty} i t_i$.

The log likelihood is expressed as

$$\begin{aligned} L &= -N_0(1-\theta)^{\gamma+1}\frac{(1-\theta)^{-\gamma}-1}{\theta\gamma} + (n-v)\log\theta + v(\gamma+1)\log(1-\theta) \\ &\quad + \sum_{i=1}^{\infty} t_i(\log\Gamma(i+\gamma) - \log\Gamma(1+\gamma)) + const. \end{aligned}$$

The first derivatives are

$$\frac{\partial L}{\partial\theta} = -\frac{N_0}{\gamma}\left(\frac{(1-\theta)^{\gamma+1}-1}{\theta^2} + \frac{(1+\gamma)(1-\theta)^{\gamma}}{\theta}\right) + (n-v)\frac{1}{\theta} - v\frac{(\gamma+1)}{1-\theta},$$

$$\begin{aligned} \frac{\partial L}{\partial\gamma} &= N_0\frac{1-\theta}{\theta\gamma^2} + \frac{N_0}{\theta}\left(-\frac{(1-\theta)^{\gamma+1}}{\gamma^2} + \frac{(1-\theta)^{\gamma+1}}{\gamma}\log(1-\theta)\right) \\ &\quad + v\log(1-\theta) + \sum_{i=2}^{\infty} t_i \sum_{j=1}^{i-1}\frac{1}{\gamma+j}. \end{aligned}$$

The ML estimators are the solution of these simultaneous equations: $\partial L/\partial\gamma = \partial L/\partial\theta = 0$. For its numerical evaluation, the Newton-Raphson method is available. The second derivatives are provided below.

$$\begin{aligned} \frac{\partial^2 L}{\partial\gamma\partial\theta} &= \frac{N_0((1-\theta)^{\gamma+1}-1)}{\theta^2\gamma^2} - \frac{N_0}{\theta^2\gamma}(1-\theta)^{\gamma+1}\log(1-\theta) \\ &\quad + \frac{N_0}{\theta\gamma^2}(1-\theta)^{\gamma} - \frac{N_0}{\theta\gamma}(1-\theta)^{\gamma}\log(1-\theta) - \frac{N_0}{\theta}(1-\theta)^{\gamma}\log(1-\theta) - \frac{v}{1-\theta}. \end{aligned}$$

$$\frac{\partial^2 L}{\partial \theta^2} = -\frac{N_0}{\gamma}\left(\frac{2-2(1-\theta)^{\gamma+1}}{\theta^3} - \frac{2(1+\gamma)(1-\theta)^{\gamma}}{\theta^2} - \frac{(1+\gamma)\gamma(1-\theta)^{\gamma-1}}{\theta}\right) - \frac{N-v}{\theta^2} - \frac{v(\gamma+1)}{(1-\theta)^2}.$$

$$\frac{\partial^2 L}{\partial \gamma^2} = \frac{N_0}{\theta}\left(\frac{-2(1-\theta)+2(1-\theta)^{\gamma+1}}{\gamma^3} - \log(1-\theta)\frac{2(1-\theta)^{\gamma+1}}{\gamma^2} + \log^2(1-\theta)\frac{(1-\theta)^{\gamma+1}}{\gamma}\right) - \sum_{i=2}^{\infty} t_i \sum_{j=1}^{i-1} \frac{1}{(j+\gamma)^2}.$$

In practice, the realized value of N (denoted by n) usually equals N_0. Then the likelihood equations reduce to the following:

$$\frac{\partial L}{\partial \theta} = 0 \Leftrightarrow v = \frac{N_0(1-(1-\theta)^{\gamma})(1-\theta)}{\gamma\theta} \tag{33}$$

and

$$\frac{\partial L}{\partial \gamma} = 0 \Leftrightarrow \frac{N_0(1-\theta)}{\gamma\theta}\left\{\frac{1-(1-\theta)^{\gamma}}{\gamma} + \log(1-\theta)\right\} + \sum_{i=2}^{\infty} t_i \sum_{j=1}^{i-1} \frac{1}{\gamma+j} = 0, \tag{34}$$

where the derivation of (34) depends on (33). It is easy to show that the first term of the left hand side of (34) is negative and the second term is positive, when $-1 < \gamma\,(\neq 0)$ and $0 < \theta < 1$. Namely,

$$\frac{N_0(1-\theta)}{\gamma\theta}\left\{\frac{1-(1-\theta)^{\gamma}}{\gamma} + \log(1-\theta)\right\} < 0 \tag{35}$$

and

$$\sum_{i=2}^{\infty} t_i \sum_{j=1}^{i-1} \frac{1}{\gamma+j} > 0. \tag{36}$$

The left hand side of (35) decreases as γ decreases. These facts suggest that the ENB model ($\gamma < 0$) fits better than the NB model ($\gamma > 0$) when small cells are dominant, i.e., the left hand side of (36) is large.

3.2 An application result

This section provides an example of fitting the ENB model. We will compare a fit of ENB(γ) by the ML estimation with a fit obtained by Engen (1974), who adopted pseudo estimation methods.

Table 1 shows the result of fitting the ENB model to insect data from Mehninick (1964). The total number of insects, N_0, was 2220. The columns titled "i" correspond to the frequency of insects of the same species, and "S_i" is the observed number of species of frequency i. We omit data of frequencies larger than 30 from Table 1; other observations were at 31, 36(2), 39, 48, 73, 76, 93, 120, 148, 201, 283 and 592. The columns titled "Engen" show a fit by Engen's pseudo moment method, which essentially uses the truncated negative binomial distribution. The parameter estimates by the pseudo moment method were $\hat{\gamma}_p = -0.366$, $\hat{\theta}_p = 0.997$; see Engen (1974) for more detail. The ML estimates of ENB(γ) are $\hat{\gamma}_m = -0.392$, $\hat{\theta}_m = 0.998$,

i	S_i	Engen	MLE	i	S_i	Engen	MLE	i	S_i	Engen	MLE
1	50	51.02	53.31	11	0	1.36	1.29	21	0	0.54	0.51
2	20	16.14	16.17	12	0	1.20	1.14	22	0	0.51	0.47
3	11	8.77	8.65	13	0	1.07	1.02	23	1	0.48	0.44
4	6	5.77	5.63	14	1	0.97	0.91	24	0	0.45	0.42
5	5	4.18	4.05	15	0	0.88	0.83	25	0	0.42	0.39
6	3	3.22	3.11	16	1	0.80	0.75	26	0	0.40	0.37
7	2	2.58	2.48	17	0	0.73	0.69	27	0	0.38	0.35
8	2	2.14	2.05	18	2	0.68	0.63	28	1	0.36	0.33
9	2	1.81	1.73	19	0	0.63	0.59	29	1	0.34	0.32
10	1	1.56	1.48	20	1	0.58	0.54	30	1	0.32	0.30

Table 1: Fits of the ENB model to insect data from Mehninick (1964)

which result in the fit shown in the columns "MLE". The ML estimation allocates slightly more proportion to small groups in this example than the pseudo moment method, but both fits seem reasonable.

Fitting ENB(γ) by the ML estimation does work. Engen had to rely on pseudo methods since the distribution of size indices was unknown. Because even the pseudo moment method requires numerical iteration, there is seemingly no reason to use pseudo methods now.

4 Concluding remarks

In diverse application fields, negative binomial distributions are often used for describing count data. However, the negative binomial distribution can hardly describe the data of many small groups, for which we could improve a fit by introducing ENB(γ).

The major way for describing counts has been to employ a truncated distribution or $\mathrm{P}(F_j|F_j \geq 1)$. Such a distribution is over positive integers, and this may remind us that generalizing distributions, denoted by $g(z)$, are also over positive integers. For instance, the ENB distribution is a truncated distribution and can be a generalizing distribution, from which the ENB model is produced. Then it is natural to ask the difference between the ENB distribution and the ENB model, or more generally, the difference between the direct use of a truncated distribution and its use as a generalizing distribution to generate the limiting distributions of size indices.

The difference is whether U is fixed or random: In the direct modeling, U has to be fixed at the observed number of nonempty cells; on the contrary U is a random variable when size indices are independently Poisson distributed. To illustrate, for fixed v, suppose that $F_j, j = 1, 2, \ldots, v$, are independently and identically subject to the ENB distribution (11). Then the joint distribution of these F_j's coincides with the conditional ENB model given $U = v$. Assuming the distribution of U enables us to estimate the increase of nonempty cells as the total number of individuals (N) grows. The practical importance of this advantage is apparent when we see the vast researches on this type of estimation surveyed by Bunge and Fitzpatrick (1993). By applying the limiting argument (8) to generalized Poisson distributions, U becomes Poisson distributed, whereby the total number of species in a population can be estimated for example. More detailed discussion on this issue will be held in the author's subsequent paper.

A generalized Poisson distribution is suitable to describe skew data that are observed in many cases; see Zipf (1949) or Mandelbrot (1983) for this empirical fact. There are ample reasons why we have shown some general properties of this class of distributions.

Acknowledgements

The author would like to express sincere thanks to Prof. M. Sibuya, Prof. A. Takemura, Prof. H. Yamato and an anonymous referee for comments that greatly improved the manuscript. This research is partly supported by the Japanese Ministry of Education, Culture, Sports, Science and Technology.

Appendix

The following notation is used henceforth:

$$\boldsymbol{F} = (F_1, F_2, \ldots, F_J) \in {\mathbb{N}_0}^J, \quad \boldsymbol{S} = (S_1, S_2, \ldots) \in {\mathbb{N}_0}^\infty,$$

$$\boldsymbol{S}(m) = (S_1, S_2, \ldots, S_m) \in {\mathbb{N}_0}^m, m \in \mathbb{N}.$$

A The infinite-dimensional multinomial distribution

The infinite-dimensional multinomial distribution has only finite sources of variation. We construct the joint distribution of an infinite series of size indices from the distribution of finite-dimensional $\boldsymbol{F}$.

Let us consider $\phi : {\mathbb{N}_0}^J \mapsto {\mathbb{N}_0}^\infty$, where

$$\phi(\boldsymbol{F}) = (\sum_{j=1}^{J} I(F_j = 1), \sum_{j=1}^{J} I(F_j = 2), \sum_{j=1}^{J} I(F_j = 3), \cdots).$$

This function ϕ assigns to each frequency vector its size indices. Hence for a proper probability mass function P,

$$1 = \sum_{\boldsymbol{F} \in {\mathbb{N}_0}^J} \mathrm{P}(\boldsymbol{F}) = \sum_{\boldsymbol{S} \in {\mathbb{N}_0}^\infty} \sum_{\boldsymbol{F} \in \{\boldsymbol{F} | \phi(\boldsymbol{F}) = \boldsymbol{S}\}} \mathrm{P}(\boldsymbol{F}). \tag{37}$$

It may be worthwhile to point out that

$$\{\phi(\boldsymbol{F}) | \boldsymbol{F} \in {\mathbb{N}_0}^J\} \subset {\mathbb{N}_0}^\infty$$

and $\mathrm{P}(\emptyset) = 0$. We write

$$\mathcal{P}(\boldsymbol{S}) = \sum_{\boldsymbol{F} \in \{\boldsymbol{F} | \phi(\boldsymbol{F}) = \boldsymbol{S}\}} \mathrm{P}(\boldsymbol{F}),$$

which is nonnegative for all $\boldsymbol{S} \in {\mathbb{N}_0}^\infty$. Note that $\mathcal{P}(\boldsymbol{S})$ is the sum of finite number of probabilities. Because (37) shows

$$1 = \sum_{\boldsymbol{S} \in {\mathbb{N}_0}^\infty} \mathcal{P}(\boldsymbol{S}),$$

we can regard $\mathcal{P}$ as the definition of the joint probability of an infinite series of size indices when the distribution of $\boldsymbol{F}$ is proper. This construction is valid for all J. In particular, the distribution of $\boldsymbol{F}$ is proper when $F_j, j \in \mathbb{N}(J)$, is independently and identically subject to a distribution over $\mathbb{N}_0$. Then the distribution of $\boldsymbol{S}$ is proper and formally specified as follows.

The restricted size indices $\boldsymbol{S}(m)$ are subject to the $m+1$ dimensional multinomial distribution p_m defined by the pgf:

$$G(z_1, z_2, \ldots, z_m) = \{\sum_{i=1}^{m}(z_i - 1)\mathrm{P}(F_1 = i) + 1\}^J, \tag{38}$$

which converges for $|z_i| \leq 1$. Then the sequence $\{p_m\}_{m=1}^{\infty}$ determines the distribution of $\boldsymbol{S}$ uniquely; see Corollary 2.20 of Breiman (1968). Let $A_m = \{\boldsymbol{S}|\boldsymbol{S}(m) = (t_1, t_2, \ldots, t_m)\}$. Since A_m are decreasing,

$$\lim_{m\to\infty} \mathrm{P}(A_m) = \mathrm{P}(\lim_{m\to\infty} A_m),$$

where $\mathrm{P}(A_m)$ is measured by p_m. We can thus write the probability mass function of $\boldsymbol{S}$ as (3).

B Proofs

Proof of Theorem 1 Given (38), the pgf of $\boldsymbol{S}(m)$ under the assumption is expressed as

$$G(z_1, z_2, \ldots, z_m) = (1 + \frac{1}{J}\mu \sum_{i=1}^{m}(z_i - 1)\mathrm{P}(F_1 = i)\frac{1}{a})^J. \tag{39}$$

In fact it is true that

$$\lim_{a\to 0} \frac{\mathrm{P}(F_1 = i)}{a} = q_i, \quad i \in \mathbb{N}. \tag{40}$$

If (40) holds then the pgf converges as $J \to \infty$ when $\mu < \infty$ and $|z_i| \leq 1$:

$$\lim_{\substack{J\to\infty \\ a\to 0}} G(z_1, z_2, \ldots, z_m) = \exp(\sum_{i=1}^{m}(z_i - 1)q_i\mu). \tag{41}$$

The right hand side of (41) implies that each S_i is independently subject to $Po(q_i\mu)$. The limiting joint distribution is proper because the right hand side of (41) equals one when $z_i = 1$ for all i. The above argument holds for all m, and thus the limiting distribution of $\boldsymbol{S}$ is determined by the sequence of the joint distribution of m independent Poisson variables as $m \to \infty$. Note that (40) is equivalent to the condition (15) of Lemma 1 when $c_i = q_i\mu$. Consequently, it suffices to prove (40).

We now show

$$\lim_{a\to 0} \sum_{i=1}^{\infty} z^i \frac{\mathrm{P}(F_1 = i)}{a} = \sum_{i=1}^{\infty} z^i q_i = g(z), \tag{42}$$

which implies (40). Because $G(z) = \sum_{i=0}^{\infty} z^i \mathrm{P}(F_1 = i)$, the left hand side of (42) equals $\lim_{a\to 0}(G(z) - G(0))/a$, which amounts to

$$\begin{aligned} &\lim_{a\to 0} \frac{\exp(a(g(z) - 1)) - \exp(a(-1))}{a} \\ &= \lim_{a\to 0}(g(z) - 1)\exp(a(g(z) - 1)) + \exp(-a) \\ &= g(z) \end{aligned}$$

by l'Hopital's rule. Q.E.D.

The proof of Theorem 2 requires the following lemma shown by Khatri and Patel (1961).

Lemma 2 *Suppose that a random variable F is subject to the generalized Poisson distribution defined by* (17). *Then*

$$(i+1)\mathrm{P}(F=i+1) = a\sum_{j=0}^{i}(i+1-j)q_{i+1-j}\mathrm{P}(F=j).$$

Proof of Theorem 2 First we show, by induction, that the assumed distribution of F_1 belongs to the class of MPS distributions. We assume (19), which is true when $i=0$ because $G(0)=\mathrm{P}(F_1=0)=1/\exp(\alpha\eta(\theta))$ and $b_0=1$. For a generalized Poisson distribution, we can use the recurrence formula stated in Lemma 2. Therefore,

$$(i+1)\mathrm{P}(F_1=i+1) = \alpha\eta(\theta)\sum_{j=0}^{i}(i+1-j)\frac{a_{i+1-j}h(\theta)^{i+1-j}}{\eta(\theta)}\mathrm{P}(F_1=j). \tag{43}$$

Using (19), we rewrite the right hand side of (43) as

$$\alpha\eta(\theta)\sum_{j=0}^{i}(i+1-j)\frac{a_{i+1-j}h(\theta)^{i+1-j}}{\eta(\theta)}\frac{b_j h(\theta)^j}{\exp(\alpha\eta(\theta))} = \frac{h(\theta)^{i+1}}{\exp(\alpha\eta(\theta))}\alpha\sum_{j=0}^{i}(i+1-j)a_{i+1-j}b_j$$

Namely,

$$\mathrm{P}(F_1=i+1) = \frac{b_{i+1}h(\theta)^{i+1}}{\exp(\alpha\eta(\theta))},$$

which again satisfies (19). Hence the distribution of F_1 belongs to the MPS class.

Let $N=\sum_{j=1}^{J}F_j$. Because the pgf of N is $G(z)^J$,

$$\mathrm{P}(N=n) = \frac{d_n h(\theta)^n}{\exp(J\alpha\eta(\theta))}.$$

Dividing

$$\mathrm{P}(F_1=g_1,F_2=g_2,\ldots,F_J=g_J,N=n) = \frac{h(\theta)^n}{\exp(J\alpha\eta(\theta))}\prod_{j=1}^{J}b_{g_j}, \quad n=\sum_{j=1}^{J}g_j, \tag{44}$$

by $\mathrm{P}(N=n)$, we obtain the conditional distribution (20).

A referee suggested to rewrite the right hand side of (20) as

$$\frac{1}{d_n\prod_{i=1}^{n}t_i!}\prod_{i=1}^{n}(Jb_i)^{t_i}(1-\frac{1}{J})\cdots(1-\frac{v-1}{J}).$$

Then, because (40) implies that $Jb_i \to a_i\mu$ as $J\to\infty$, the last result of the theorem obviously holds. The author originally obtained the limiting distribution by conditioning the joint distribution of infinite Poisson variables. Q.E.D.

Proof of Theorem 3 We first consider the limiting of $\alpha \to 0$. Jørgensen (1982, p.171) stated that for $\gamma > 0$

$$\alpha^{\gamma} K_{\gamma}(\alpha) \to \Gamma(\gamma) 2^{\gamma - 1} \tag{45}$$

as $\alpha \to 0$. Because, as seen in Jørgensen (1982, p.170) for instance,

$$K_{\gamma}(\alpha) = K_{-\gamma}(\alpha)$$

we obtain

$$\alpha^{-\gamma} K_{\gamma}(\alpha) \to \Gamma(-\gamma) 2^{-\gamma - 1} \tag{46}$$

as $\alpha \to 0$ when $\gamma < 0$.

According to Sichel (1974), $\mathrm{E}(N) = N_0$ is equivalent to

$$N_0 = J \frac{\alpha\theta}{2\sqrt{1-\theta}} \frac{K_{\gamma+1}(\alpha\sqrt{1-\theta})}{K_{\gamma}(\alpha\sqrt{1-\theta})} \tag{47}$$

under (32). Since $-1 < \gamma < 0$, the restriction (47) becomes

$$N_0 = (\frac{2}{\alpha})^{2\gamma} \frac{J\theta\Gamma(\gamma+1)}{\Gamma(-\gamma)(1-\theta)^{\gamma+1}} \tag{48}$$

as $\alpha \to 0$ by (45) and (46). The right hand side of (48) is constant if and only if $J\alpha^{-2\gamma}$ is fixed as $\alpha \to 0$ and $J \to \infty$. Therefore we consider the limiting argument of

$$J\alpha^{-2\gamma} = \mu \text{ fixed}, \quad \text{as } J \to \infty \text{ and } \alpha \to 0 \tag{49}$$

for $-1 < \gamma < 0$, where

$$\mu = \frac{N_0 \Gamma(-\gamma)(1-\theta)^{\gamma+1}}{2^{2\gamma}\theta\Gamma(\gamma+1)}.$$

By Lemma 1, it suffices to show under (49) that

$$\lim_{J\to\infty} J\mathrm{P}(F = i) = \mu \frac{\theta^i \Gamma(i+\gamma) 2^{2\gamma}}{\Gamma(-\gamma) i!}, \tag{50}$$

where $\mathrm{P}(F = i)$ is given in (31); see the proof of Theorem 1. We can easily show (50) for $-1 < \gamma < 0$, considering (45) and (46). Q.E.D.

References

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

[2] Breiman, L. (1968). *Probability*. Addison-Wesley.

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

[4] Charalambides, C.A. and Singh, J. (1988). A review of the Stirling numbers, their generalizations and statistical applications. *Communications in Statistics*, Theor. Meth., **17**, 2533–2595.

[5] Engen, S. (1974). On species frequency models. *Biometrika*, **61**, 263–270.

[6] Engen, S. (1977). Comments on two different approaches to the analysis of species frequency data. *Biometrics*, **33**, 205–213.

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

[8] Ewens, W.J. (1972). The sampling theory of selectively neutral alleles. *Theoretical Population Biology*, **3**, 87–112.

[9] Feller, W. (1957). *An Introduction to Probability Theory and its Applications, Vol. 1*, 2nd, Wiley, New York.

[10] 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.

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

[12] Gupta, R.C. (1974). Modified power series distributions and some of its applications. *Sankhyā*, B, **36**, 288–298.

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

[14] Hoshino, N. (2002). 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.

[15] Hoshino, N. (2003). Random clustering based on the conditional inverse Gaussian-Poisson distribution. *Journal of the Japan Statistical Society*, **33**, 105–117.

[16] 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.

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

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

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

[20] Kemp, A.W. (1978). Cluster size probabilities for generalized Poisson distributions. *Communications in Statistics*, Theor. Meth., **7**, 1433–1438.

[21] Khatri, C.G. and Patel, I.R. (1961). Three classes of univariate discrete distributions. *Biometrics*, **17**, 567–575.

[22] Mandelbrot, B.B. (1983). *The Fractal Geometry of Nature*. W. H. Freeman and Company, New York.

[23] Mehninick, E.F. (1964). A comparison of some species individuals diversity indices applied to samples of field insects. *Ecology*, **45**, 859–861.

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

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

[26] Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields*, **102**, 145–158.

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

[28] Sibuya, M. (1979). Generalized hypergeometric, digamma and trigamma distribution. *Annals of Institute of Statistical Mathematics*, **31**, 373–390.

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

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

[31] 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.), S.A. C.S.I.R., Pretoria, 51–97.

[32] Sichel, H.S. (1974). On a distribution representing sentence-length in written prose. *Journal of the Royal Statistical Society*, A, **137**, 25–34.

[33] Sichel, H.S. (1992). Anatomy of the generalized inverse Gaussian-Poisson distribution with special applications to bibliometric studies. *Information Processing and Management*, **28**, 5–17.

[34] 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.

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

[36] Watterson, G.A. (1974). Models for the logarithmic species abundance distributions. *Theoretical Population Biology*, **6**, 217–250.

[37] Willmot, G.E. (1986). Mixed compound Poisson distributions. *ASTIN Bulletin*, **16**, S59–S79.

[38] Willmot, G.E. (1989). A remark on the Poisson-Pascal and some other contagious distributions. *Statistics and Probability Letters*, **7**, 217–220.

[39] Yamato, H., Sibuya, M. and Nomachi, T. (2001). Ordered sample from two-parameter GEM distribution. *Statistics and Probability Letters*, **55**, 19–27.

[40] Zipf, G.K. (1949). *Human Behavior and the Principle of Least Effort.* Addison-Wesley, Cambridge, MA.