Ann Inst Stat Math (2012) 64:457–474
DOI 10.1007/s10463-011-0327-8

Random partitioning over a sparse contingency table

Nobuaki Hoshino

Received: 3 February 2010 / Revised: 2 December 2010 / Published online: 6 March 2011

Abstract The present article investigates a class of random partitioning distributions of a positive integer. This class is called the limiting conditional compound poisson (LCCP) distribution and characterized by the law of small numbers. Accordingly the LCCP distribution explains the limiting behavior of counts on a sparse contingency table by the frequencies of frequencies. The LCCP distribution is constructed via some combinations of conditioning and limiting, and this view reveals that the LCCP distribution is a subclass of several known classes that depend on a Bell polynomial. It follows that the limiting behavior of a Bell polynomial provides new asymptotics for a sparse contingency table. Also the Neyman Type A distribution and the Thomas distribution are revisited as the basis of the sparsity.

Keywords Discrete multivariate distribution · Infinitely divisible · Size index · Statistical disclosure control · Species abundance

1 Introduction

A sparse contingency table implies that a sample size n is far smaller than the number of cells J. This situation arises, e.g., from a case–control study of a rare disease, which involves hundreds of variables for only a few hundred samples. A standard practice for data of this kind avoids cross-classifying with respect to all variables; a table with fewer cells is constructed for fewer samples. This dependence of J on n leads to the standard sparse asymptotics that n/J converges to a positive constant as n and J go

N. Hoshino (✉)
School of Economics, Kanazawa University, Kakuma-machi,
Kanazawa-shi, Ishikawa, 920-1192, Japan
e-mail: hoshino@kenroku.kanazawa-u.ac.jp

to infinity (see, e.g. Fienberg and Holland 1973). By taking $n \to \infty$, central limit theorems apply.

The present article substitutes the law of small numbers for the central limit theorem. More precisely, we fix n while $J \to \infty$. The Poisson distribution governs this limit, where n individuals are randomly partitioned. A class of these partitioning distributions is studied in the present article.

We will take $n \to \infty$ for this class, which provides alternative sparse asymptotics. The proposed limiting argument is motivated by practices in which J is very large regardless of n. For example, in ecology, let J be the number of species, which may include extinct species. Then J does not apparently depend on the number of observed individuals n. In statistical disclosure control the risk of breaching privacy is evaluated for a data set similar to the rare disease example. This kind of risk is considered as a function of the frequencies of cells, and is assessed with respect to J for fixed n (see Hoshino 2009).

The objective class of random partitioning is called the limiting conditional compound poisson (LCCP) distribution by Hoshino (2009), since its derivation employs the limiting and conditioning of compound Poisson distributions. An equivalent of the LCCP distribution is introduced as a discretization of an infinitely divisible distribution over nonnegative real numbers in Hoshino (2006).

The present article shows that the LCCP distribution is more generally derived by the law of small numbers. Also clarified is the relationship among the LCCP distribution and other classes of random partitioning distributions. It turns out that the law of small numbers characterizes the LCCP distribution, which is thus worth consideration in particular.

The following subsection introduces more detailed contexts and the developments of the present article.

1.1 Setup

Throughout the present article, $\mathbb{N}_0$ and $\mathbb{N}$ are, respectively, the sets of nonnegative integers and positive integers. For $n \in \mathbb{N}$, $[n] := \{1, 2, \ldots, n\}$.

In our modeling of a contingency table, the frequency of the jth cell is denoted by $F_{j,J}$, $j \in [J]$. The sum of the frequencies is

$$N_J := \sum_{j=1}^{J} F_{j,J}. \tag{1}$$

A standard model of a contingency table supposes that $F_{j,J}$, $j \in [J]$, is independently distributed over $\mathbb{N}_0$. For example, the joint distribution of $\boldsymbol{F}_J := (F_{1,J}, F_{2,J}, \ldots, F_{J,J})$ is often the product of the Poisson distribution with mean λ_j, $j \in [J]$, which is denoted by $\mathtt{Po}(\lambda_j)$ henceforth. We express the independence of random variables by "$\times$" such as $\boldsymbol{F}_J \sim \times_{j=1}^{J} \mathtt{Po}(\lambda_j)$, where "$\sim$" implies "is distributed as". The conditional distribution of this $\boldsymbol{F}_J$ given $N_J = n$ is multinomial, which is frequently used as well. The following argument does not assume a specific distribution for $F_{j,J}$, but they are assumed to be independent before conditioning on $N_J = n$.

To describe a sparse contingency table, we will take $J \to \infty$ while $N_J = n$ is fixed. Since almost every cell is empty in the limit, we consider the limiting distribution of

$$S_{i,J} := \sum_{j=1}^{J} I(F_{j,J} = i), \quad i \in \mathbb{N}. \tag{2}$$

These statistics are known as frequencies of frequencies (Good 1953) or size indices (Sibuya 1993). It holds $N_J = \sum_{i=1}^{\infty} i S_{i,J}$. Rather the behavior of size indices than that of cell frequencies is sometimes of interest in practice such as statistical disclosure control.

If the limiting distribution of a random variable X is the same as that of another random variable Y, we write $X \xrightarrow{d} Y$. In this article $S_{i,J} \xrightarrow{d} S_i$ and $N_J \xrightarrow{d} N$ as $J \to \infty$, where

$$S_i \sim \mathtt{Po}(\mu q_i), \quad i \in \mathbb{N}, \tag{3}$$

independently. That is,

$$\boldsymbol{S} := (S_1, S_2, \ldots) \sim \times_{i=1}^{\infty} \mathtt{Po}(\mu q_i). \tag{4}$$

We may shortly write $S_{i,J} \xrightarrow{d} \mathtt{Po}(\mu q_i)$. Our canonical expression requires

$$0 < \mu < \infty, \quad q_i \geq 0,\ i \in \mathbb{N}, \quad \sum_{i=1}^{\infty} q_i = 1. \tag{5}$$

Then $\boldsymbol{q} := \{q_i\}_{i \in \mathbb{N}}$ is a proper distribution over $\mathbb{N}$. It can be shown that $N := \sum_{i=1}^{\infty} i S_i$ has the compound Poisson distribution, which is defined for some proper $\boldsymbol{q}$ by the following probability generating function (pgf):

$$G(z) = \exp(\mu(g(z) - 1)), \quad \mu > 0, \tag{6}$$

where

$$g(z) = \sum_{i=1}^{\infty} z^i q_i. \tag{7}$$

We refer to (6) by $\mathtt{CP}(\mu, \boldsymbol{q})$ (see Johnson et al. 1993, p. 188 for more on this distribution).

Example Assume that $g(z)$ is of the logarithmic series distribution whose probability mass function (pmf) is defined for $0 < \theta < 1$ as $\left\{\frac{1}{-\log(1-\theta)} \frac{\theta^i}{i}\right\}_{i \in \mathbb{N}} =: \mathtt{LS}(\theta)$.

Then $g(z) = \log(1-\theta z)/\log(1-\theta)$, and $\mathtt{CP}(-k\log(1-\theta), \mathtt{LS}(\theta))$ is the negative binomial distribution, whose pgf is $G(z) = \exp(-k\log(1-\theta)(\log(1-\theta z)/\log(1-\theta)-1))$. □

Arratia et al. (2003) point out that many random combinatorial structures can be regarded as the conditional distribution of independent random variables $X_1, X_2, \ldots, X_n$ given $\sum iX_i = n$. If X_i is Poisson distributed, the resulting conditional distribution is called assembly. Hoshino (2006, 2009) derives this class by limiting and conditioning of the compound Poisson distribution; this derivation will be reviewed in Proposition 1. We formally introduce the class of interest below.

Definition 1 Suppose that (4) holds. Then, for $n \in \mathbb{N}$, we call the conditional distribution of $\boldsymbol{S}$ given $\sum_{i=1}^{\infty} iS_i = n$ the LCCP distribution generated by $\boldsymbol{q}$. We refer to this distribution by $\mathtt{LCCP}(\mu, \boldsymbol{q})$.

If $N = n$, then S_i has to be zero for all $i > n$. Therefore we regard the LCCP distribution as the distribution of n dimensional vector $\boldsymbol{S}_n := (S_1, \ldots, S_n)$. The support of an LCCP distribution is $\mathcal{S}_{|n} := \left\{\boldsymbol{s}_n : s_i \in \mathbb{N}_0, i \in [n], \sum_{i=1}^{n} is_i = n\right\}$, where $\boldsymbol{s}_n := (s_1, s_2, \ldots, s_n)$. This set $\mathcal{S}_{|n}$ coincides with the set of all unordered partitions of a positive integer n. Hence the LCCP distribution can be interpreted as a class of random partitioning distributions.

Three examples of the LCCP distribution are presented below. The most famous one is called the Ewens (1972) distribution, which is surveyed by Johnson et al. (1997, Chap. 41) and more closely to our context by Charalambides (2007, Sec. 4.2). The limiting conditional inverse-Gaussian Poisson (LCIGP) distribution is proposed by Hoshino (2006). The limiting quasi-multinomial (LQM) distribution derived by Hoshino (2005b) is equivalent to Moon's model of Pitman (1999). For an LCCP distribution, we denote $u := \sum_{i=1}^{n} s_i$.

Example (*Ewens*) For $k > 0$, $\mathtt{LCCP}(-k\log(1-\theta), \mathtt{LS}(\theta))$ is called the Ewens distribution, whose pmf is

$$\left\{\frac{k^u n!}{k(k+1)\cdots(k+n-1)}\prod_{i=1}^{n}\left(\frac{1}{i}\right)^{s_i}\frac{1}{s_i!}\right\}_{\boldsymbol{s}_n\in\mathcal{S}_{|n}} =: \mathtt{Ew}(k).$$

□

Example (*LCIGP*) Engen (1974) proposes the extended (truncated) negative binomial distribution, whose special case has $\boldsymbol{q}$ of $\left\{\frac{1}{1-\sqrt{1-\theta}}\frac{\theta^i(2i-3)!!}{2^i i!}\right\}_{i\in\mathbb{N}} =: \mathtt{ENB}(\theta)$, $0 < \theta \leq 1$. For $\mu > 0$, $\mathtt{LCCP}(\mu(1-\sqrt{1-\theta}), \mathtt{ENB}(\theta))$ is called the LCIGP distribution, whose pmf is

$$\left\{\sqrt{\frac{\pi}{2\mu}}\frac{n!\exp(-\mu)}{\mu^{n-u}K_{n-1/2}(\mu)}\prod_{i=1}^{n}\left(\frac{(2i-3)!!}{i!}\right)^{s_i}\frac{1}{s_i!}\right\}_{\boldsymbol{s}_n\in\mathcal{S}_{|n}} =: \mathtt{LCIGP}(\mu),$$

where $K_{n-1/2}(\cdot)$ is the modified Bessel function of the third kind of order $n-1/2$. □

Example (*LQM*) The Borel (1942) distribution has the pmf of the following: $\left\{\frac{(\lambda i)^{i-1}}{i!}\exp(-\lambda i)\right\}_{i\in\mathbb{N}} =: \mathtt{Bo}(\lambda), 0 < \lambda \leq 1$. For $\rho > 0$, $\mathtt{LCCP}(\rho\lambda, \mathtt{Bo}(\lambda))$ is called the LQM distribution. Its pmf is

$$\left\{n!\,\rho^{u-1}(\rho+n)^{1-n}\prod_{i=1}^{n}\left(\frac{i^{i-1}}{i!}\right)^{s_i}\frac{1}{s_i!}\right\}_{\boldsymbol{s}_n\in\mathcal{S}_n} =: \mathtt{LQM}(\rho).$$

□

Hoshino's (2009, Theorem 4) derivation of the LCCP distribution reads:

Proposition 1 *Suppose that*

$$\boldsymbol{F}_J \sim \times_{j=1}^{J}\mathtt{CP}(\lambda_j, \boldsymbol{q}), \quad \lambda_j \geq 0,\ j \in [J]. \tag{8}$$

Then two conditions

$$\lim_{J\to\infty}\sum_{j=1}^{J}\lambda_j = \mu, \quad 0 < \mu < \infty, \tag{9}$$

and

$$\lim_{J\to\infty}\max_{j}\lambda_j = 0 \tag{10}$$

are sufficient for

$$((S_{1,J}, S_{2,J}, \ldots | N_J), N_J) \xrightarrow{d} (\mathtt{LCCP}(\mu, \boldsymbol{q}), \mathtt{CP}(\mu, \boldsymbol{q})) \tag{11}$$

as $J \to \infty$.

We will see that the limiting arguments (9) and (10) comprise the law of small numbers. Figure 1 depicts the idea of Proposition 1; the LCCP distribution can be derived from (8) in two ways by changing the order of the limiting and the conditioning.

Section 2 shows that the law of small numbers is necessary and sufficient for (11) in a broader sense. Namely the law of small numbers characterizes the LCCP distribution among the class of random partitioning distributions. Considering other classes of random partitioning, we note the construction of Kolchin model (Kerov 1995) is partly the same as Fig. 1. Hence Sect. 2 also explicates the construction of the LCCP distribution to include this literature. In Sect. 3 we take $n \to \infty$ for alternative sparse asymptotics. Some results in terms of a Bell polynomial are stated for the LCCP distribution. In Sect. 4, the Neyman Type A distribution and the Thomas distribution exemplify the argument of the present article. Appendix A notes that Karlin (1967) model of size indices is slightly different from ours. Appendix B gathers the proofs of theorems.

$$\begin{tikzcd}
\boxed{(S_{1,J}, S_{2,J}, \ldots \mid N_J)} \arrow[d, "J\to\infty"'] & \boxed{\begin{array}{c}(S_{1,J}, S_{2,J}, \ldots)\\ (2)\end{array}} \arrow[l, "conditioning"', "N_J"] \arrow[d, "J\to\infty"] \\
\boxed{\begin{array}{c}(S_1, S_2, \ldots \mid N) \sim \mathtt{LCCP}(\mu, \boldsymbol{q})\\ \text{Definition 1}\end{array}} & \boxed{\begin{array}{c}(S_1, S_2, \ldots) \sim \times_{i=1}^{\infty} \mathtt{Po}(\mu q_i)\\ (4)\end{array}} \arrow[l, "conditioning"', "N\sim\mathtt{CP}(\mu,\boldsymbol{q})"]
\end{tikzcd}$$

Fig. 1 The law of small numbers

2 The construction of the LCCP distribution

This section clarifies that the LCCP distribution can be derived in a broader situation than that of Proposition 1. Section 2.1 proves that size indices converge to the LCCP distribution if and only if the law of small numbers holds. Section 2.2 generalizes Fig. 1 construction of the LCCP distribution by conditioning on the number of non-empty cells. Section 2.3 considers the special case of compound Poisson frequencies to understand the implication of the generalized construction. In Sect. 2.4, the LCCP distribution connects to Kolchin's class of random partitioning distributions by the generalized construction.

2.1 The characterization of the LCCP distribution

To show that the law of small numbers is equivalent to the size indices' convergence to the LCCP distribution, we employ Koopman (1950) necessary and sufficient condition for (3):

Proposition 2 *Suppose that $F_{j,J}$, $j \in [J]$, is independently distributed over $\mathbb{N}_0$. Then* (3) *holds if and only if*

$$\lim_{J\to\infty} \max_{j} \mathrm{P}(F_{j,J} = i) = 0, \quad i \in \mathbb{N}, \tag{12}$$

and

$$\lim_{J\to\infty} \mathrm{E}(S_{i,J}) = \mu q_i, \quad i \in \mathbb{N}. \tag{13}$$

The limiting argument of (12) and (13) is essentially the law of small numbers. This fact is clear when we regard a size index $S_{i,J}$ as the number of successes of Bernoulli trials with unequal success probabilities. This type of distribution is called Poisson's binomial by Wang (1993) (see also Johnson et al. 1993, p. 138).

Wang and Ji (1993, Theorem 2) show that (12) and (13) are equivalent to that $N = \sum_{i=1}^{\infty} i S_i \sim \mathtt{CP}(\mu, \boldsymbol{q})$. Therefore, as $J \to \infty$,

$$(S_{1,J}, S_{2,J}, \ldots, N_J) \xrightarrow{d} (\boldsymbol{S}, N) \sim (\times_{i=1}^{\infty} \mathtt{Po}(\mu q_i), \mathtt{CP}(\mu, \boldsymbol{q})) \tag{14}$$

is equivalent to the law of small numbers. Rewriting (14), we have the following theorem.

Theorem 1 *Let $F_{j,J}$, $j \in [J]$, be independently distributed over $\mathbb{N}_0$. Then (11) holds as $J \to \infty$ if and only if the law of small numbers (12) and (13) hold.*

Remark 1 The compound Poisson distribution is equivalent to the infinitely divisible distribution over $\mathbb{N}_0$ (see, e.g. Steutel and van Harn 2004, Theorem 3.2, p. 30). An infinitely divisible distribution is equivalent to the limiting sum of uniformly almost negligible random variables (see, e.g. Steutel and van Harn 2004, Theorem 5.3, p. 15). Hence Wang and Ji's (1993, Theorem 2) result is a discrete special case.

It is important that Theorem 1 does not assume the distribution of $\boldsymbol{F}_J$. Moreover, (14) still holds for "weakly" dependent $F_{j,J}$'s (see, e.g. Meyer 1973). Therefore the LCCP distribution describes sparse contingency tables in many situations.

Proposition 1 deals with a special case of Theorem 1; we can show the following corollary. For a different example of the limiting argument of Theorem 1, see Hoshino (2005a, Theorem 2.3).

Theorem 2 *In Proposition 1, the two sufficient conditions (9) and (10) are also necessary.*

2.2 Conditioning on the number of nonempty cells

Theorem 1 implies that Fig. 1 is valid for $F_{j,J}, j \in [J]$, that is independently distributed over $\mathbb{N}_0$. This subsection further expands the idea of Fig. 1 by conditioning on the number of nonempty cells. For later use this subsection requires general notation.

The pmf of a size indices vector $(S_{1,J}, S_{2,J}, \ldots)$ is denoted by

$$\pi_J(\boldsymbol{s}_n) = \mathrm{P}((S_{1,J}, \ldots, S_{N_J,J}) = \boldsymbol{s}_n),$$
$$\boldsymbol{s}_n \in \mathcal{S}_J := \cup_{n=0}^{\infty} \left\{ \boldsymbol{s}_n : s_i \in \mathbb{N}_0, i = 1, 2, \ldots, n, \sum_{i=1}^{n} i s_i = n, \sum_{i=1}^{n} s_i \leq J \right\}. \quad (15)$$

It is noteworthy that N_J may be 0, and we treat $\boldsymbol{s}_0$ as empty. The conditional distribution of π_J given $N_J = n$ has the pmf of

$$\pi_{J|n}(\boldsymbol{s}_n) = \frac{\pi_J(\boldsymbol{s}_n)}{\sum_{\boldsymbol{s}_n \in \mathcal{S}_{J|n}} \pi_J(\boldsymbol{s}_n)},$$
$$\boldsymbol{s}_n \in \mathcal{S}_{J|n} := \left\{ \boldsymbol{s}_n : s_i \in \mathbb{N}_0, i \in [n], \sum_{i=1}^{n} i s_i = n, \sum_{i=1}^{n} s_i \leq J \right\}. \quad (16)$$

The pmf of the limiting distribution of π_J as $J \to \infty$ is

$$\pi(\boldsymbol{s}_n; \mu, \boldsymbol{q}) = \prod_{i=1}^{n} \frac{\exp(-\mu q_i)(\mu q_i)^{s_i}}{s_i!}, \quad \boldsymbol{s}_n \in \cup_{n=0}^{\infty} \mathcal{S}_{|n} =: \mathcal{S}. \quad (17)$$

$$\begin{tikzcd}
\begin{array}{c}\pi_{J|n}, \mathcal{S}_{J|n} \\ (16)\end{array} \arrow[d, "limit{:}(12,13)"] & \begin{array}{c}\pi_{J}, \mathcal{S}_{J} \\ (15)\end{array} \arrow[l, "conditioning"', "N_J"] \arrow[r, "conditioning", "U_J\sim\texttt{PoBin}"'] \arrow[d, "limit{:}(12,13)"] & \begin{array}{c}\pi_{J|u}, \mathcal{S}_{J|u} \\ (20)\end{array} \arrow[d, "limit{:}(12,13)"] \\
\begin{array}{c}\pi_{|n}, \mathcal{S}_{|n} \\ (18)\end{array} & \begin{array}{c}\pi, \mathcal{S} \\ (17)\end{array} \arrow[l, "conditioning"', "N\sim\texttt{CP}(\mu,\boldsymbol{q})"] \arrow[r, "conditioning", "U\sim\texttt{Po}(\mu)"'] & \begin{array}{c}\pi_{|u}, \mathcal{S}_{|u} \\ (21)\end{array}
\end{tikzcd}$$

Fig. 2 Relationship among size indices' distributions

We denote the pmf of $\texttt{LCCP}(\mu, \boldsymbol{q})$ by

$$\pi_{|n}(\boldsymbol{s}_n; \mu, \boldsymbol{q}) = \frac{\pi(\boldsymbol{s}_n)}{\sum_{\boldsymbol{s}_n \in \mathcal{S}_{|n}} \pi(\boldsymbol{s}_n)}, \quad \boldsymbol{s}_n \in \mathcal{S}_{|n}. \tag{18}$$

Next let us consider the number of nonempty cells denoted by

$$U_J := \sum_{j=1}^{J} I(F_{j,J} \geq 1) = \sum_{i=1}^{\infty} S_{i,J} \tag{19}$$

or $U := \sum_{i=1}^{\infty} S_i \sim \texttt{Po}(\mu)$. We observe that U_J has Poisson's binomial distribution with success probabilities $\mathrm{P}(F_{j,J} \geq 1)$, $j \in [J]$. This distribution converges in distribution to $\texttt{Po}(\mu)$ by the law of small numbers: (12) and (13).

We express the conditional distributions of π_J and π given $U_J = u$ and $U = u$ by $\pi_{J|u}$ and $\pi_{|u}$. Namely

$$\pi_{J|u}(\boldsymbol{s}_n) = \frac{\pi_J(\boldsymbol{s}_n)}{\sum_{\boldsymbol{s}_n \in \mathcal{S}_{J|u}} \pi_J(\boldsymbol{s}_n)},$$
$$\boldsymbol{s}_n \in \mathcal{S}_{J|u} := \cup_{n=u}^{\infty} \left\{ \boldsymbol{s}_n : s_i \in \mathbb{N}_0, i \in [n], \sum_{i=1}^{n} i s_i = n, \sum_{i=1}^{n} s_i = u \leq J \right\}, \tag{20}$$

and

$$\pi_{|u}(\boldsymbol{s}_n) = \frac{\pi(\boldsymbol{s}_n)}{\sum_{\boldsymbol{s}_n \in \mathcal{S}_{|u}} \pi(\boldsymbol{s}_n)},$$
$$\boldsymbol{s}_n \in \mathcal{S}_{|u} := \cup_{n=u}^{\infty} \left\{ \boldsymbol{s}_n : s_i \in \mathbb{N}_0, i \in [n], \sum_{i=1}^{n} i s_i = n, \sum_{i=1}^{n} s_i = u \right\}. \tag{21}$$

Since $\lim_{J\to\infty} \mathcal{S}_{J|u} = \mathcal{S}_{|u}$ and $\pi_J \to \pi$, we have the result below generalizing Fig. 1 to Fig. 2.

Theorem 3 *Suppose that $F_{j,J}$, $j \in [J]$, is independently distributed over $\mathbb{N}_0$. Then as we apply the law of small numbers* (12) *and* (13),

$$\lim_{J\to\infty} \pi_{J|u}(\boldsymbol{s}_n) = \pi_{|u}(\boldsymbol{s}_n), \quad \boldsymbol{s}_n \in \mathcal{S}_{|u}. \tag{22}$$

The right hand side of (22) can be explicitly written as

$$\pi_{|u}(\boldsymbol{s}_n; \boldsymbol{q}) = \frac{u!}{s_1! \cdots s_n!} \prod_{i=1}^{n} {q_i}^{s_i}, \quad \boldsymbol{s}_n \in \mathcal{S}_{|u}, \tag{23}$$

which is multinomial. We note that $\pi_{|u}$ does not depend on μ, or U is sufficient for μ of π. Size indices are multinomially distributed when frequencies are independent and identically distributed (see, e.g. Hoshino 2005a, Appendix A). This fact is specifically stated below.

Proposition 3 *Let $X_1, \ldots, X_u$ be independent and identically distributed as $\boldsymbol{q}$. Denote a size index by $T_i = \sum_{j=1}^{u} I(X_j = i), i \in \mathbb{N}$. When $n = \sum_{j=1}^{u} x_j$ is the observed sum of frequencies,* $\mathrm{P}((T_1, \ldots, T_n) = \boldsymbol{s}_n) = \pi_{|u}(\boldsymbol{s}_n; \boldsymbol{q})$.

2.3 An example of compound Poisson frequencies

This subsection considers a special case of Fig. 2 where frequencies are independently compound Poisson distributed. By this example we can well understand the role of the law of small numbers. $\mathtt{CP}$ is a practical class of distributions over $\mathbb{N}_0$ since it over-disperses (see Johnson et al. 1993, p. 354). Hoshino (2009) validates $\mathtt{CP}$ in modeling a contingency table, for it is closed under the corruption of cells.

Throughout this subsection we employ the assumption of Proposition 1 or (8). Then Hoshino (2009) calls $\pi_{J|n}$ the conditional compound poisson (CCP) distribution generated by $\boldsymbol{q}$; Theorem 2 states that (9) and (10) are necessary and sufficient for the CCP distribution's convergence to the LCCP distribution. Actually $N_J \sim \mathtt{CP}\left(\sum_{j=1}^{J} \lambda_j, \boldsymbol{q}\right)$. Hence (9) alone is equivalent to $N \sim \mathtt{CP}(\mu, \boldsymbol{q})$. To understand the meaning of (10), let us denote the pgf of $\mathtt{CP}(\lambda_j, \boldsymbol{q})$ by $G_j(z)$. Then

$$\mathrm{P}(F_{j,J} = 0) = G_j(0) = \exp(-\lambda_j). \tag{24}$$

Therefore in considering the definition (19) of U_J, (10) implies that the success probability of the Poisson's binomial distribution goes to zero. Consequently U_J converges to the Poisson distribution.

Simultaneously zero truncated distribution of $F_{j,J}$ converges to $\boldsymbol{q}$. Let $\tilde{F}_{j,J}$ be zero-truncated $F_{j,J}$, or $\mathrm{P}(\tilde{F}_{j,J} = i) := \mathrm{P}(F_{j,J} = i)/(1 - \mathrm{P}(F_{j,J} = 0)) = \mathrm{P}(F_{j,J} = i | F_{j,J} \geq 1)$, $i \in \mathbb{N}$. Then Kemp (1978) shows

$$\lim_{\lambda_j \to 0} \mathrm{P}(\tilde{F}_{j,J} = i) = q_i, \quad i \in \mathbb{N}. \tag{25}$$

$$\boxed{\begin{array}{c}\varpi_{|n} \text{ over } \mathcal{S}_{|n}\\ (27)\end{array}} \xleftarrow[N]{\textit{conditioning}} \boxed{\begin{array}{c}\varpi \text{ over } \mathcal{S}\\ (26)\end{array}} \xleftarrow[U\sim\boldsymbol{v}]{\textit{mixture}} \boxed{\begin{array}{c}\pi_{|u} \text{ over } \mathcal{S}_{|u}\\ (23)\end{array}}$$

Fig. 3 Kolchin's modeling

This result confirms that $\pi_{J|u} \to \pi_{|u}$ under (8), which was suggested by Professor Akimichi Takemura. Zero truncation is equivalent to conditioning on nonempty cells, and by the law of small numbers, all the positive frequencies are i.i.d. as $\boldsymbol{q}$ in the limit.

2.4 Kolchin's model

Generalizing the idea of Kolchin (1971), Kerov (1995) formulates a class of random partitioning distributions called Kolchin's model, which contains the class of LCCP distributions. To see this fact, let us review the construction of a Kolchin model.

For $u \in \mathbb{N}_0$, suppose that u random variables are i.i.d. as $\boldsymbol{q}$. Then the size indices of these are multinomially distributed as (23) or $\pi_{|u}$. Take a distribution over $\mathbb{N}_0$ of $\boldsymbol{v} := \{v_u\}_{u=0}^{\infty}$, with which we mix $\pi_{|u}$ as

$$\sum_{u=0}^{\infty} v_u \cdot \pi_{|u}(\boldsymbol{s}_n; \boldsymbol{q}) =: \varpi(\boldsymbol{s}_n; \boldsymbol{v}, \boldsymbol{q}), \quad \boldsymbol{s}_n \in \mathcal{S}. \tag{26}$$

Conditioning ϖ on n, we have a random partitioning distribution:

$$\varpi_{|n}(\boldsymbol{s}_n; \boldsymbol{v}, \boldsymbol{q}) := \frac{\varpi(\boldsymbol{s}_n; \boldsymbol{v}, \boldsymbol{q})}{\sum_{\boldsymbol{s}_n \in \mathcal{S}_{|n}} \varpi(\boldsymbol{s}_n; \boldsymbol{v}, \boldsymbol{q})}, \quad \boldsymbol{s}_n \in \mathcal{S}_{|n}. \tag{27}$$

This construction is illustrated as Fig. 3.

Definition 2 The distribution of (27) is called Kolchin's model with parameters $\boldsymbol{v}$ and $\boldsymbol{q}$.

It is obvious from Fig. 2 that

$$\pi(\boldsymbol{s}_n; \mu, \boldsymbol{q}) = \varpi(\boldsymbol{s}_n; \mathtt{Po}(\mu), \boldsymbol{q}). \tag{28}$$

An immediate proposition follows:

Proposition 4 *Kolchin's model with parameters* $\mathtt{Po}(\mu)$ *and* $\boldsymbol{q}$ *equals the LCCP distribution generated by* $\boldsymbol{q}$. *Equivalently,*

$$\pi_{|n}(\boldsymbol{s}_n; \mu, \boldsymbol{q}) = \varpi_{|n}(\boldsymbol{s}_n; \mathtt{Po}(\mu), \boldsymbol{q}). \tag{29}$$

Consequently an LCCP distribution has the property of a Kolchin model. Using this result, we can show the uniqueness of the Ewens distribution among LCCP distributions on Kingman (1978) partition structure defined below.

Definition 3 Let $p_n(\cdot)$ be some pmf over $\mathcal{S}_{|n}$. If for all $n \in \mathbb{N}$, $p_n(\boldsymbol{s}_n) = p_{n+1}(s_1+1, s_2, \ldots, s_{n+1})\frac{s_1+1}{n+1} + \sum_{r=2}^{n+1} p_{n+1}(s_1, \ldots, s_{r-1}-1, s_r+1, \ldots, s_{n+1})\frac{r(s_r+1)}{n+1}$, then the distribution of $p_n(\cdot)$ is said to have partition structure.

Definition 3 implies that a given partition of n elements results from the deletion of one element uniformly at random from a partition of $n+1$ elements. This property thus assures that a model is closed under simple random sampling without replacement.

Theorem 4 *Among LCCP distributions, only the Ewens distribution has partition structure.*

3 Using Bell polynomials

The LCCP distribution can be expressed as an expansion of a Bell polynomial. Pitman (2006) formulates this expression as the Gibbs partition, which is named after statistical physics (see Vershik's 1996 explanation).

Based on this expression, we take $n \to \infty$ for the LCCP distribution. Then dependence among size indices should diminish since conditioning on $N = n$ becomes less restrictive. Hence we expect a size index S_i converges to independent Poisson; this surmise is formalized together with other consequences in this section.

First let us define a (total) Bell polynomial denoted by $B_n(x_1, \ldots, x_n) := n! \sum_{\boldsymbol{s}_n \in \mathcal{S}_{|n}} \prod_{i=1}^n (x_i/i!)^{s_i}/s_i!$. A partial Bell polynomial is defined by $B_{n,u}(x_1, \ldots, x_n) := n! \sum_{\boldsymbol{s}_n \in \mathcal{S}_{|n,u}} \prod_{i=1}^n (x_i/i!)^{s_i}/s_i!$, where $\mathcal{S}_{|n,u} := \{\boldsymbol{s}_n : s_i \in \mathbb{N}_0, i \in [n], \sum_{i=1}^n i s_i = n, \sum_{i=1}^n s_i = u\}$. It follows

$$\sum_{u=1}^{n} \mu^u B_{n,u}(x_1, \ldots, x_n) = B_n(\mu x_1, \ldots, \mu x_n) \tag{30}$$

(see, e.g. Charalambides 2002, eq. 11.15).

The pgf (6) of $\mathtt{CP}(\mu, \boldsymbol{q})$ multiplied by $\exp(\mu)$ is the generating function of Bell polynomials. When $N \sim \mathtt{CP}(\mu, \boldsymbol{q})$,

$$\mathrm{P}(N = n) = \frac{\exp(-\mu)}{n!} B_n(\mu x_1, \ldots, \mu x_n). \tag{31}$$

Therefore a Bell polynomial inevitably appears when we deal with a compound Poisson distribution.

The Gibbs partition uses the weighted sum of partial Bell polynomials as a normalizing constant. Write $B_n(\boldsymbol{w}, \boldsymbol{x}) := \sum_{u=1}^n w_u B_{n,u}(\boldsymbol{x})$, where $\boldsymbol{w} = (w_1, w_2, \ldots)$ and $\boldsymbol{x} = (x_1, x_2, \ldots)$.

Definition 4 A Gibbs partition with parameters $\boldsymbol{w}, \boldsymbol{x}$ is defined by the following pmf:

$$\gamma_{|n}(\boldsymbol{s}_n; \boldsymbol{w}, \boldsymbol{x}) := \frac{n! w_u}{B_n(\boldsymbol{w}, \boldsymbol{x})} \prod_{i=1}^{n} \left(\frac{x_i}{i!}\right)^{s_i} \frac{1}{s_i!}, \quad \boldsymbol{s}_n \in \mathcal{S}_{|n}.$$

Because of (30) a Gibbs partition is not uniquely determined by the parameters. For example,

$$\gamma_{|n}(\boldsymbol{s}_n; (1, 1, \ldots), (\mu x_1, \mu x_2, \ldots)) = \gamma_{|n}(\boldsymbol{s}_n; (\mu, \mu^2, \ldots), (x_1, x_2, \ldots)). \tag{32}$$

Pitman (2006, Theorem 1.2) points out that a Gibbs partition has a representation of Kolchin model. In particular, we have the following expression.

Proposition 5 *An LCCP distribution is a Gibbs partition of* (32). *That is, for* $n \in \mathbb{N}$,

$$\pi_{|n}(\boldsymbol{s}_n; \mu, \boldsymbol{q}) = \frac{n!\mu^u \prod_{i=1}^n {q_i}^{s_i} \frac{1}{s_i!}}{B_n(\mu x_1, \ldots, \mu x_n)}, \quad \boldsymbol{s}_n \in \mathcal{S}_{|n}, \tag{33}$$

where $x_i = i!\, q_i$, $u = \sum_{i=1}^n s_i$.

Next we consider a special case where $\boldsymbol{q}$ belongs to the class of power series distributions. Then $\boldsymbol{q}$ is expressed for a power parameter $\theta > 0$ as

$$\left\{ \frac{y_i \theta^i}{\eta(\theta) i!} \right\}_{i \in \mathbb{N}} =: \mathtt{PS}(\boldsymbol{y}, \theta). \tag{34}$$

In this case, x_i in (33) equals $y_i \theta^i / \eta(\theta)$. By (see, e.g. Charalambides 2002, eq. 11.3)

$$B_n(\theta^1 y_1, \theta^2 y_2, \ldots, \theta^n y_n) = \theta^n B_n(y_1, y_2, \ldots, y_n) \tag{35}$$

and (30), we have

$$\pi_{|n}(\boldsymbol{s}_n; \mu\eta(\theta), \mathtt{PS}(\boldsymbol{y}, \theta)) = \frac{n!\mu^u \prod_{i=1}^n \left(\frac{y_i}{i!}\right)^{s_i} \frac{1}{s_i!}}{B_n(\mu y_1, \ldots, \mu y_n)}, \quad \boldsymbol{s}_n \in \mathcal{S}_{|n}. \tag{36}$$

Remark 2 The right hand side of (36) does not depend on the power parameter θ. Figure 2 explains this fact as that, in Proposition 3, $\sum_{i=1}^u X_i$ or N is sufficient for θ.

Example (*Ewens*) Suppose that $y_i = (i-1)!$ and $\eta(\theta) = -\log(1-\theta)$. Then $\mathtt{LS}(\theta) = \mathtt{PS}(\boldsymbol{y}, \theta)$. Let $\mu = k$, and the denominator in (36) reduces to

$$B_n(k0!, k1!, \ldots, k(n-1)!) = k(k+1)\cdots(k+n-1) \tag{37}$$

(see, e.g. Charalambides 2002, eq. 8.4). Consequently $\pi_{|n}(\boldsymbol{s}_n; -k\log(1-\theta), \mathtt{LS}(\theta)) = \mathtt{Ew}(k)$. □

In the following we study a random vector of size indices $\boldsymbol{S}_n \sim \mathtt{LCCP}(\mu, \boldsymbol{q})$. The marginal moments are cited from Hoshino (2009):

Proposition 6 *Suppose that* $\boldsymbol{S}_n \sim \texttt{LCCP}(\mu, \boldsymbol{q})$. *Then for all* $r_1, \ldots, r_n \in \mathbb{N}_0$ *such that* $l := \sum_{i=1}^n i r_i \leq n$, *the factorial moments are*

$$\mathrm{E}\left(\prod_{i=1}^n S_i^{(r_i)}\right) = \frac{B_{n-l}(\mu x_1, \ldots, \mu x_{n-l})\mu^r n^{(l)}}{B_n(\mu x_1, \ldots, \mu x_n)} \prod_{i=1}^n \left(\frac{x_i}{i!}\right)^{r_i}, \tag{38}$$

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

Sibuya (1993) takes $n \to \infty$ for the Ewens distribution and shows that the first m components of $\boldsymbol{S}_n$ converge to independent Poisson distributions. Similar results for LCIGP and LQM are shown by Hoshino (2006, 2005b). An analogue for a general LCCP distribution is given below.

Theorem 5 *Suppose that* $\boldsymbol{S}_n \sim \texttt{LCCP}(\mu, \boldsymbol{q})$. *Let* m *be a finite fixed positive integer. If and only if*

$$\lim_{n\to\infty} \frac{n B_{n-1}(\mu x_1, \ldots, \mu x_{n-1})}{B_n(\mu x_1, \ldots, \mu x_n)} = c < \infty, \tag{39}$$

the first m *components* $(S_1, S_2, \ldots, S_m) = \boldsymbol{S}_m$ *converge as* $n \to \infty$ *to* $\times_{i=1}^m \texttt{Po}(c^i \mu q_i)$.

It must be $c \geq 1$ in (39) if we require $\texttt{CP}(\mu, \boldsymbol{q})$ to be proper. If $\sum_{n=0}^{\infty} \mathrm{P}(N = n) = 1$, d'Alembert's ratio test concludes

$$\lim_{n\to\infty} \frac{\mathrm{P}(N = n-1)}{\mathrm{P}(N = n)} \geq 1. \tag{40}$$

By (31) the left hand side of (40) equals the left hand side of (39).

Example (*Ewens*) We examine the condition (39) on $\texttt{Ew}(k)$ using (37): $\lim_{n\to\infty} nk(k+1)\cdots(k+n-2)/(k(k+1)\cdots(k+n-1)) = 1$. Therefore $\boldsymbol{S}_m$ converges to independent $\texttt{Po}(k/i)$, $i = 1, 2, \ldots, m$. □

To see that the LCCP distribution belongs to an exponential family, we rewrite the pmf as $\pi_{|n}(\boldsymbol{s}_n; \mu, \boldsymbol{q}) = \exp\left(u \log \mu + \log n! - \log B_n(\mu x_1, \ldots, \mu x_n)\right) \prod_{i=1}^n {q_i}^{s_i}/s_i!$. Regarding $\log \mu$ as the unique parameter, we have the following statement (see Theorem 5.6 of Lehmann 1991).

Theorem 6 *Suppose that* $\boldsymbol{S}_n \sim \texttt{LCCP}(\mu, \boldsymbol{q})$. *Then it belongs to an exponential family, and*

$$U_n := \sum_{i=1}^n S_i \tag{41}$$

is complete and sufficient for μ.

The sufficient statistic U_n is of interest in many applications such as the abundance of species (see a survey on this statistic by Bunge and Fitzpatrick 1993). If $\boldsymbol{S}_n \sim \mathtt{LCCP}(\mu, \boldsymbol{q})$, the distribution of U_n is

$$\mathrm{P}(U_n = u) = \frac{B_{n,u}(\mu x_1, \ldots \mu x_n)}{B_n(\mu x_1, \ldots, \mu x_n)}, \quad u \in [n]. \tag{42}$$

Nandi and Dutta (1988) consider the special case of (42) where $\boldsymbol{q} = \mathtt{PS}$. We treat the right hand side of (42) as a class of distributions:

Definition 5 The generalized Bell distribution generated by $\boldsymbol{x}$ is defined for $\mu > 0$ by the pmf of (42), which is referred to by $\mathtt{GB}(\mu, \boldsymbol{x})$.

Differently from our definition, in Nandi and Dutta (1988), the generalized Bell distribution refers to the case where the denominator is Enneking and Ahuja's (1976) generalized Bell number.

Uppuluri and Carpenter (1969) discuss the moment properties of $\mathtt{GB}(1, (1, 1, \ldots))$ or the Bell distribution. Its pmf reduces to $S(n, u)/B_n(1, 1, \ldots), u \in [n]$, where $S(n, u) = B_{n,u}(1, 1, \ldots)$ is the Stirling number of the second kind. The denominator $B_n(1, 1, \ldots)$ is the Bell number (see, e.g. Riordan 1968, p. 192).

Next we consider the limiting distribution of U_n as $n \to \infty$. Pitman (2006, p. 33) reviews the study of a central limit theorem for U_n, i.e., $(U_n - \mathrm{E}(U_n))/\sqrt{\mathrm{V}(U_n)}$ converges in distribution to the standard normal. We alternatively generalize the cases of $\mathtt{LQM}(\mu)$ and $\mathtt{LCIGP}(\mu)$, for which $U_n \xrightarrow{d} 1 + \mathtt{Po}(\mu)$ is shown by Hoshino (2005b, 2006). Interestingly the Ewens distribution has a different limiting distribution of U_n (see, e.g. Arratia et al. 2003, Section 4.2). This difference can be explained by the asymptotic expression of a partial Bell polynomial.

Theorem 7 *Let $f(\cdot)$ be some function and c be a positive finite real number. If*

$$B_{n,u}(x_1, \ldots, x_n) \approx \frac{c^u}{(u-1)!} f(n), \quad u \in [n], \tag{43}$$

when n is large, then $\mathtt{GB}(\mu, \boldsymbol{x})$ converges in distribution to $1 + \mathtt{Po}(c\mu)$ as $n \to \infty$.

It is worthy of note that $f(n)$ must not depend on u in (43).

Example (*LCIGP*) For $\mathtt{LCIGP}(\mu)$, $x_i = (2i-3)!!/2^i, i \in \mathbb{N}$. Then $B_n(\mu x_1, \ldots, \mu x_n) = \sqrt{\frac{2\mu}{\pi}} \frac{K_{n-1/2}(\mu)}{\exp(-\mu)} \left(\frac{\mu}{2}\right)^n$, and $B_{n,u}(x_1, \ldots, x_n) = \frac{(2n-u-1)!}{(u-1)!(n-u)!} \left(\frac{1}{2}\right)^{2n-u}$. Using Stirling's formula of $n! \approx \sqrt{2\pi}(n/e)^n$, we have $B_{n,u}(x_1, \ldots, x_n) \approx \frac{n^{n-1}}{e^n\sqrt{2}} \frac{1}{(u-1)!}$, which is the case of $c = 1$, $f(n) = n^{n-1}/(e^n\sqrt{2})$ in (43). □

4 Two cases for a sparse contingency table

This section investigates two more examples of the LCCP distribution. Recalling that $\boldsymbol{q}$ is the limiting distribution of the zero-truncated frequency of a cell, we may assume that $\boldsymbol{q}$ is the Poisson distribution since it could be the frequency of a rare event. However, we require $\boldsymbol{q}$ be distributed over $\mathbb{N}$, and two modifications to the Poisson distribution are considered. The first one is zero truncation, which follows $\boldsymbol{q} = \left\{\frac{\phi^i \exp(-\phi)}{i!} \frac{1}{1-\exp(-\phi)}\right\}_{i\in\mathbb{N}} =: \mathtt{TPo}(\phi)$. The second one is shifting, which follows $\boldsymbol{q} = \left\{\frac{\phi^{i-1}\exp(-\phi)}{(i-1)!}\right\}_{i\in\mathbb{N}} = 1 + \mathtt{Po}(\phi)$. Below we observe that these two cases result in different LCCP distributions.

The pgf of $\mathtt{TPo}(\phi)$ is expressed as $g(z) = (\exp(\phi z) - 1)/(\exp(\phi) - 1)$. Therefore the pgf of $\mathtt{CP}(\mu, \mathtt{TPo}(\phi))$ is

$$G(z) = \exp\left[\mu\left(\frac{\exp(\phi z) - 1}{\exp(\phi) - 1} - 1\right)\right] = \exp\left[\frac{\mu\exp(\phi)}{\exp(\phi) - 1}\left(\frac{\exp(\phi z)}{\exp(\phi)} - 1\right)\right]. \quad (44)$$

The last expression of (44) implies that $\mathtt{Po}(\phi)$ is compounded as $\boldsymbol{q}$. This is the usual form of the Neyman Type A distribution, which is reviewed by Johnson et al. (1993, p. 368). Thus $\mathtt{CP}(\mu, \mathtt{TPo}(\phi))$ is the Neyman Type A distribution.

$\mathtt{TPo}(\phi)$ is $\mathtt{PS}(\boldsymbol{y}, \phi)$, where $\eta(\phi) = \exp(-\phi)/(1 - \exp(-\phi))$, $y_i = 1, i \in \mathbb{N}$. Using (36), $\mathtt{LCCP}\left(\frac{\mu\exp(-\phi)}{1-\exp(-\phi)}, \mathtt{TPo}(\phi)\right)$ has the following pmf:

$$\pi_{|n}\left(\boldsymbol{s}_n; \frac{\mu\exp(-\phi)}{1 - \exp(-\phi)}, \mathtt{TPo}(\phi)\right) = \frac{n!\mu^u \prod_{i=1}^n \left(\frac{1}{i!}\right)^{s_i} \frac{1}{s_i!}}{B_n(\mu, \ldots, \mu)}, \quad \boldsymbol{s}_n \in \mathcal{S}_{|n}. \quad (45)$$

Equation (45) does not depend on the power parameter ϕ as noted in Remark 2. We also observe that $\mathtt{LCCP}\left(\frac{\mu\exp(-\phi)}{1-\exp(-\phi)}, \mathtt{TPo}(\phi)\right)$ belongs to an exponential family with one parameter μ as in Theorem 6. Its complete sufficient statistic U_n is distributed as $\mathtt{GB}(\mu, (1, 1, \ldots))$. Reminding of $B_{n,u}(\mu, \ldots, \mu) = \mu^u S(n, u)$, $\mathtt{GB}(\mu, (1, 1, \ldots))$ is regarded as a power-series-distributionized Bell distribution. Since (see, e.g. Charalambides 2002, p. 323) $S(n, u) \approx u^n/u!$ when n is large, Theorem 7 suggests that $\mathtt{GB}(\mu, (1, 1, \ldots))$ does not converge to $1 + \mathtt{Po}(\mu)$.

Next we consider the other case; the pgf of $1 + \mathtt{Po}(\phi)$ is $g(z) = z\exp(\phi(z - 1))$. The resulting compound Poisson distribution $\mathtt{CP}(\mu, 1 + \mathtt{Po}(\phi))$ is called the Thomas (1949) distribution (see Johnson et al. 1993, p. 392).

Again $1 + \mathtt{Po}(\phi)$ is $\mathtt{PS}(\boldsymbol{y}, \phi)$, where $\eta(\phi) = \exp(-\phi)/\phi$, $y_i = i, i \in \mathbb{N}$. Using (36), $\mathtt{LCCP}(\mu\exp(-\phi)/\phi, 1 + \mathtt{Po}(\phi))$ has the following pmf:

$$\pi_{|n}\left(\boldsymbol{s}_n; \frac{\mu\exp(-\phi)}{\phi}, 1 + \mathtt{Po}(\phi)\right) = \frac{n!\mu^u \prod_{i=1}^n \left(\frac{1}{(i-1)!}\right)^{s_i} \frac{1}{s_i!}}{B_n(\mu, 2\mu, 3\mu, \ldots, n\mu)}, \quad \boldsymbol{s}_n \in \mathcal{S}_{|n}. \quad (46)$$

Equation (46) does not depend on the power parameter ϕ either. Consequently $\mathtt{LCCP}(\mu\exp(-\phi)/\phi, 1+\mathtt{Po}(\phi))$ belongs to an exponential family with one parameter μ, whose sufficient statistic U_n is distributed as $\mathtt{GB}(\mu, (1, 2, 3, \ldots, n))$. Comtet (1974, p. 135) calls $B_{n,u}(1, 2, 3, \ldots, n)$ the idempotent number, and $B_{n,u}(\mu, 2\mu, \ldots, n\mu) = \mu^u\binom{n}{u}u^{n-u}$. Since $B_{n,u}(1, 2, 3, \ldots, n) \approx n^u u^{n-u}/u!$, Theorem 7 suggests that $\mathtt{GB}(\mu, (1, 2, 3, \ldots, n))$ does not converge to $1 + \mathtt{Po}(\mu)$.

These two LCCP distributions are applicable to sparse contingency table analysis with the general results provided in Sect. 3.

Appendix

A. Karlin's model

The framework of the LCCP distribution is similar to Karlin's model but different in the following sense. Karlin (1967) considers an urn model where n balls are thrown independently at a fixed infinite array of cells with probability q_i of hitting the ith cell. Let $X_{n,i}$ be the number of balls in the ith cell after n tosses. If n is subject to a Poisson process $\{N(t); t \in [0, \infty)\}$ with parameter 1, $X_{N(t),i}$ or the number of balls in the ith cell at time t is independently Poisson distributed with parameter tq_i. Namely $\mathrm{P}(X_{N(\mu),1} = s_1, X_{N(\mu),2} = s_2, \ldots) = \pi(s_1, s_2, \ldots; \mu, \boldsymbol{q})$. The right hand side is our model of size indices $\boldsymbol{S} := (S_1, S_2, \ldots)$. Confusingly, Karlin (1967) is interested in the distribution of size indices: $Z_r(t) := \sum_{i=1}^{\infty} I(X_{N(t),i} = r), r \in \mathbb{N}_0$. Hence $Z_r(\mu) \overset{d}{=} \sum_{i=1}^{\infty} I(S_i = r)$, which is the size index of size indices in our sense. Also we should note that $N(\mu) \overset{d}{=} U = \sum_{i=1}^{\infty} S_i \neq N = \sum_{i=1}^{\infty} iS_i$. The same distribution is used for different concepts.

B. Proofs

Proof of Theorem 2 We will show (9) and (10) are equivalent to the two conditions of Theorem 1: (12) and (13).

We rewrite $\mathrm{E}(S_{i,J}) = \sum_{j=1}^{J} \mathrm{P}(F_{j,J} = i) = \sum_{j=1}^{J} \lambda_j \mathrm{P}(F_{j,J} = i)/\lambda_j$. Hence (13) is equivalent to

$$\lim_{J\to\infty} \sum_{j=1}^{J} \lambda_j \frac{\mathrm{P}(F_{j,J} = i)}{\lambda_j} = \mu q_i, \quad i \in \mathbb{N}. \tag{47}$$

Hoshino (2005a, eq. B.2) shows that $\lim_{\lambda_j \to 0} \mathrm{P}(F_{j,J} = i)/\lambda_j = q_i, i \in \mathbb{N}$, when (8) holds. Therefore (9) and (10) implies (47) or (13). Also (10) implies (12) because of (24).

On the contrary, if (12) holds then $\lim_{J\to\infty} \mathrm{P}(F_{j,J} = 0) = 1$ for all j. This is equivalent to (10) because of (24). Hence by (12), (13) reduces to $\lim_{J\to\infty} \sum_{j=1}^{J} \lambda_j q_i = \mu q_i, i \in \mathbb{N}$. Therefore (12) and (13) imply (9) and (10). □

Proof of Theorem 4 This result is rather immediate from Kerov (1995, Theorem 7.1), who shows that when Kolchin's model has partition structure then for $\boldsymbol{s}_n \in \mathcal{S}_{|n}$

$$\varpi_{|n}(\boldsymbol{s}_n; \boldsymbol{v}, \boldsymbol{q}) = n! \frac{\theta^{[u:\alpha]}}{\theta^{[n]}} \prod_{j=1}^{n} \left(\frac{(1-\alpha)^{[j-1]}}{j!} \right)^{s_j} \frac{1}{s_j!} =: \mathtt{Pit}(\alpha, \theta), \tag{48}$$

where $\theta^{[u:\alpha]} = \theta(\theta+\alpha)\cdots(\theta+(u-1)\alpha)$, $\theta^{[n]} = \theta(\theta+1)\cdots(\theta+n-1)$. The right hand side of (48) defines Pitman's (1995) distribution, whose parameter space includes limits.

The LCCP distribution requires that $\boldsymbol{v} = \mathtt{Po}(\mu)$. Then in Proposition 6.3 of Kerov (1995), $y = 0$, $b = \mu > 0$ and α becomes zero. Since $\mathtt{Pit}(0, \theta) = \mathtt{Ew}(\theta)$, an LCCP distribution that has partition structure has to be $\mathtt{Ew}(\theta)$. □

Proof of Theorem 5 We show the result by the method of moments (see, e.g. Breiman 1992, p. 181). If for all $r_1, \ldots, r_m \in \mathbb{N}_0$

$$\lim_{n\to\infty} \mathrm{E}\left(\prod_{i=1}^{m} S_i^{(r_i)}\right) = \prod_{i=1}^{m} (c^i \mu q_i)^{r_i}, \tag{49}$$

then $\boldsymbol{S}_m \xrightarrow{d} \times_{i=1}^{m} \mathtt{Po}(c^i \mu q_i)$. Conversely if $\boldsymbol{S}_m \xrightarrow{d} \times_{i=1}^{m} \mathtt{Po}(c^i \mu q_i)$ then (49) holds. Therefore we show the equivalence of (49) to (39).

By (38),

$$\mathrm{E}\left(\prod_{i=1}^{m} S_i^{(r_i)}\right) = \frac{B_{n-\ell}(\mu x_1, \ldots, \mu x_{n-\ell}) n!}{B_n(\mu x_1, \ldots, \mu x_n)(n-\ell)!} \prod_{i=1}^{m} (\mu q_i)^{r_i}, \tag{50}$$

where $\ell = \sum_{i=1}^{m} i r_i$. Thus (49) is tantamount to

$$\lim_{n\to\infty} \prod_{j=0}^{\ell-1} \frac{(n-j) B_{n-j-1}(\mu x_1, \ldots, \mu x_{n-j-1})}{B_{n-j}(\mu x_1, \ldots, \mu x_{n-j})} = c^{\ell}. \tag{51}$$

If (39) holds then (51) holds for all $r_1, \ldots, r_m$. Conversely (51) reduces to (39) when $\ell = 1$. Hence the equivalence has been proved. □

Proof of Theorem 7 We denote the pgf of $\mathtt{GB}(\mu, \boldsymbol{x})$ by $G_{\mathrm{GB}}(z) = \sum_{u=1}^{n} z^u B_{n,u}(\mu x_1, \ldots, \mu x_n)/B_n(\mu x_1, \ldots, \mu x_n)$. It suffices to show that $\lim_{n\to\infty} G_{\mathrm{GB}}(z) = z\exp(c\mu(z-1))$, which is the pgf of the shifted Poisson distribution.

If the condition (43) holds, $B_n(zx_1, \ldots, zx_n) = \sum_{u=1}^{n} z^u B_{n,u}(x_1, \ldots, x_n) \approx \sum_{u=1}^{n} \frac{(cz)^u}{(u-1)!} f(n) \to cz\exp(cz) f(n)$ as $n \to \infty$. Then $G_{\mathrm{GB}}(z) \to \{c\mu z \exp(c\mu z) f(n)\}/\{c\mu \exp(c\mu) f(n)\} = z\exp(c\mu(z-1))$. □

Acknowledgments This research has been supported by Grant-in-Aid for Scientific Research awarded by Japan Society for the Promotion of Science. Prof. Masaaki Sibuya, Prof. Akimichi Takemura and Prof. Hajime Yamato suggested many improvements on the subject of the present paper. The author would like to thank them.

References

Arratia, R., Barbour, A. D., Tavarè, S. (2003). *Logarithmic combinatorial structures: A probabilistic approach*. Zürich: European Mathematical Society

Borel, E. (1942). Sur l'emploi du théorème de Bernoulli pour faciliter le calcul d'un infinité de coefficients. Application au problème de l'attente à un guichet. *Comptes Rendus, Académie des Sciences, Paris, Series A, 214*, 452–456

Breiman, L. (1992). *Probability*. Philadelphia: SIAM

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

Charalambides, Ch. A. (2002). *Enumerative combinatorics*. Florida: Chapman and Hall/CRC

Charalambides, Ch. A. (2007). Distributions of random partitions and their applications. *Methodology and Computing in Applied Probability, 9*, 163–193

Comtet, L. (1974). *Advanced combinatorics*. Boston: D. Reidel Publishing Company

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

Fienberg, S. E., Holland, P. W. (1973). Simultaneous estimation of multinomial cell probabilities. *Journal of the American Statistical Association, 68*, 683–691

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

Hoshino, N. (2005a). Engen's extended negative binomial model revisited. *Annals of the Institute of Statistical Mathematics, 57*, 369–387

Hoshino, N. (2005b). *On a limiting quasi-multinomial distribution*. Discussion Paper CIRJE-F-361. Tokyo: Faculty of Economics, University of Tokyo

Hoshino, N. (2006). A discrete multivariate distribution resulting from the law of small numbers. *Journal of Applied Probability, 43*, 852–866

Hoshino, N. (2009). The quasi-multinomial distribution as a tool for disclosure risk assessment. *Journal of Official Statistics, 25*, 269–291

Johnson, N. L., Kotz, S., Kemp, A. W. (1993). *Univariate discrete distributions* (2nd ed.). New York: Wiley

Karlin, S. (1967). Central limit theorems for certain infinite urn schemes. *Journal of Mathematics and Mechanics, 17*, 373–401

Kerov, S. (1995). *Coherent random allocations and the Ewens–Pitman formula. PDMI Preprint*. St. Petersburg: Steklov Math Institute

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

Kolchin, V. F. (1971). A problem on allocations of particles into boxes, and the cycles of random permutations. *Teoriya Veroyatnostei i ee Primeneniya, 16*, 67–82

Koopman, B. O. (1950). Necessary and sufficient conditions for Poisson's distribution. *Proceedings of the American Mathematical Society, 1*, 813–823

Lehmann, E. L. (1991). *Theory of point estimation*. California: Wadsworth

Meyer, R. M. (1973). A Poisson-type limit theorem for mixing sequences of dependent 'rare' events. *Annals of Probability, 1*, 480–483

Nandi, S. B., Dutta, S. K. (1988). Some developments in the generalized Bell distribution. *Sankhyā, B, 50*, 362–375

Pitman, J. (1999). Coalescent random forests. *Journal of Combinatorial Theory, A, 85*, 165–193

Pitman, J. (2006). *Combinatorial stochastic processes. Lecture notes in mathematics* (Vol. 1875). New York: Springer

Riordan, J. (1968). *Combinatorial identities*. New York: Wiley

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

Steutel, F. W., van Harn, K. (2004). *Infinite divisibility of probability distributions on the real line*. New York: Marcel Dekker

Thomas, M. (1949). A generalization of Poisson's binomial limit for use in ecology. *Biometrika, 36*, 18–25

Uppuluri, V. R., Carpenter, J. R. (1969). Numbers generated by the function $\exp(1 - e^x)$. *The Fibonacci Quarterly, 7*, 437–448

Vershik, A. M. (1996). Statistical mechanics of combinatorial partitions, and their limit shapes. *Functional Analysis and its Applications, 30*, 90–105

Wang, Y. H. (1993). On the number of successes in independent trials. *Statistica Sinica, 3*, 295–312

Wang, Y. H., Ji, S. (1993). Derivations of the compound Poisson distribution and process. *Statistics and Probability Letters, 18*, 1–7