Записки научных
семинаров ПОМИ
Том 325, 2005 г.

S. Kerov

COHERENT RANDOM ALLOCATIONS, AND THE EWENS–PITMAN FORMULA[1]

ABSTRACT. Assume that there is a random number K of positive integer random variables $S_1, \dots, S_K$ which are conditionally independent, given K, and all have identical distributions. A random integer partition $N = S_1+S_2+\dots+S_K$ arises, and we denote by P_N the conditional distribution of this partition, for a fixed value of N. We prove that the distributions $\{P_N\}_{N=1}^{\infty}$ form a partition structure in the sense of Kingman if, and only if, they are governed by the Ewens–Pitman Formula. The latter generalizes the celebrated Ewens Sampling Formula which has numerous applications in pure and applied mathematics.

The distributions of random variables K and S_j belong to a family of integer distributions with two real parameters, which we call *quasi-binomial*. Hence, every Ewens–Pitman distribution arises as a result of a two-stage random procedure based on this simple class of integer distributions.

0. Introduction. In this paper, we give a simple new characterization of the Ewens–Pitman random partitions, introduced in [16]. The distribution we are interested in, $P_N^{(\alpha,\theta)}$, is defined on the set $\mathcal{P}_N$ of all partitions $N = R_1+2R_2+3R_3+\dots$ of a natural number N; it depends on two real parameters α, θ (see the equation (1.3) below). In the particular case of $\alpha = 0, \theta > 0$, the probabilities $P_N^{(\alpha,\theta)}$ are given by the *Ewens Sampling Formula*

$$P_N^{(0,\theta)}(1^{R_1}2^{R_2}\dots) = \frac{N!}{\theta(\theta+1)\dots(\theta+N-1)}\prod_{j\geqslant 1}\frac{1}{R_j!}\Big(\frac{\theta}{j}\Big)^{R_j},$$

which was a subject of intensive study in the last two decades, and has applications in pure mathematics, physics, biology, etc. (see [9, 15, 19] and also [11, 2, 4]).

[1]This paper is an edited and actualised version of the unpublished PDMI preprint 21/1995. Further development of the ideas of this work is found in [21, 25]. A number of detected misprints was fixed without notice, the bibliography extended beyond the original 19 references and a few comments added as footnotes. (Comments by *Alexander Gnedin.*)

For two consecutive values of N, the distributions $P_N = P_N^{(\alpha,\theta)}$ are related by the formula

$$\sum_{j=1}^{N+1} \frac{j(R_j+1)}{N+1} P_{N+1}(\dots,(j-1)^{R_{j-1}-1}, j^{R_j+1},\dots) \\ = P_N(1^{R_1}, 2^{R_2}, \dots, N^{R_N}).$$

We shall say that a family of distributions $\{P_N\}$ with this property is *coherent* (in terms of [13] it forms a *partition structure*).

Following an idea introduced by Kolchin in [5], we consider a rather general model of random allocations of particles into boxes. In this model, the number K of boxes is random, and obeys a given integer distribution $\varkappa$ with probabilities

$$\mathrm{Prob}\{K=k\} = \varkappa_k; \qquad k=1,2,\dots.$$

Denote by $S_1, S_2, \dots, S_K$ the random numbers of particles in each of K boxes. We assume that conditionally, given the number of boxes K, the variables $S_1, S_2, \dots, S_K$ are independent and all have the same common distribution σ:

$$\mathrm{Prob}\{S_j=k\} = \sigma_k; \qquad k=1,2,\dots.$$

Let R_j denote the number of boxes with exactly j particles, so that $N = R_1 + 2R_2 + \dots$ is a random partition. We denote by P_N the conditional distribution on the subset $\mathcal{P}_N$ of partitions with a fixed number of particles N.

Our first observation is that *one can choose the integer distributions $\varkappa, \sigma$ in Kolchin's model in such a way that $P_N = P_N^{(\alpha,\theta)}$ will coincide with the Ewens–Pitman distribution.* The second observation is as follows. *Provided the family $\{P_N\}_{N=1}^{\infty}$ of conditional distributions in Kolchin's model is coherent, we derive that $P_N = P_N^{(\alpha,\theta)}$ for appropriate values of parameters α, θ* (i.e., the family coincides with the one given by the Ewens–Pitman Formula). This characterization (Theorem 7.1) is the main result of the present paper.

We describe now the composition of the paper. In Section 1 we introduce the Ewens–Pitman Formula (1.3). In Section 2, a simple urn model leading to this formula is indicated. In Section 3, we prove the well known fact that the Ewens–Pitman distributions are coherent. The

Kolchin model is described in detail in Section 4. In Section 5 we introduce and classify a family of integer distributions which we call *quasi-binomial.* These very distributions are used in Section 6 as the parameter distributions in Kolchin's model, in order to obtain the Ewens–Pitman distributions. Our main result, Theorem 7.1, is proved in Section 7.

The author thanks A. M. Vershik, M. I. Gordin, G. I. Ol'shanskii and N. Tsilevich for useful discussions on the subject of the paper, and J. Pitman for the possibility to use his inspiring preprints [16, 18].

1. The Ewens–Pitman Formula. A *partition* of a natural number N can be given by any array of integer multiplicities $\Lambda = (R_1, R_2, \dots)$, such that

$$N = R_1 + 2R_2 + 3R_3 + \dots . \tag{1.1}$$

The component R_j is called the *multiplicity* of an integer j in the partition Λ. It follows from (1.1), that $R_j = 0$ for all $j > N$. We denote by

$$K = R_1 + R_2 + R_3 + \dots \tag{1.2}$$

the total *number of components* (or *blocks*) of the partition Λ.

The set of partitions of N is denoted by $\mathcal{Y}_N$, and by $\mathcal{Y} = \bigcup_{N=0}^{\infty} \mathcal{Y}_N$ we denote the set of all partitions. The partitions are commonly recorded in the form $\Lambda = (1^{R_1}\, 2^{R_2}\, 3^{R_3} \dots)$ thus emphasizing that R_j is the multiplicity of a component j.

In this paper, we study the *Ewens–Pitman Sampling Formula*

$$P_N^{(\alpha,\theta)}(\Lambda) = \frac{d_\Lambda}{[\theta]_N} \prod_{j=0}^{K-1} (\theta + j\,\alpha) \prod_{S \geqslant 2} \prod_{j=1}^{S-1} \Big(1 - \frac{\alpha}{j}\Big)^{R_S}; \quad \Lambda \in \mathcal{Y}_N, \tag{1.3}$$

introduced by J. Pitman in [16]. Here and in the sequel $[t]_N = t(t + 1) \dots (t + N - 1)$ is the Pochhammer symbol, and

$$d_\Lambda = \frac{N!}{1^{R_1} R_1!\, 2^{R_2} R_2! \dots} \tag{1.4}$$

for the number of permutations with the cycle structure Λ. For appropriate parameters α, θ, the equation (1.3) provides a probability distribution on the set of partitions $\mathcal{Y}_N$. The principal range of parameters is

$$0 \leqslant \alpha < 1;\ \ \theta > -\alpha \qquad \text{or} \qquad \alpha < 0;\ \ \theta/\alpha = -1, -2, -3, \dots . \tag{1.5}$$

Taking into account the limiting cases listed below, the space of admissible parameters (α, θ) looks like the one in Figure 1.

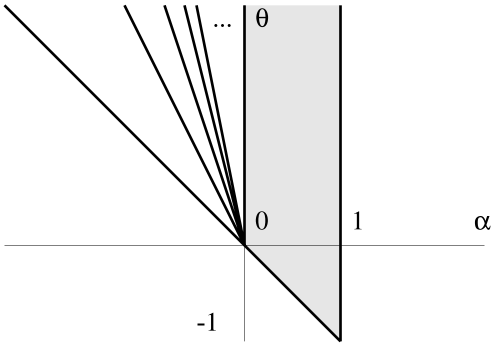

Fig. 1. The parameter space for the Ewens–Pitman Formula.

If $\alpha = 0$, the equation (1.3) takes the form

$$P_N^{(\theta)}(\Lambda) = \frac{d_\Lambda}{[\theta]_N}\,\theta^K; \qquad \Lambda \in \mathcal{Y}_N \tag{1.6}$$

and hence coincides with the Ewens Sampling Formula introduced in [8]. The distribution (1.6) was intensively studied within last two decades from various viewpoints (such as combinatorics of permutations, genetics, ecology and so on). For the contemporary survey of these applications see [9, 19, 11][2].

We list now a few degenerate limiting cases of the Ewens–Pitman Formula.

(1.7) If $\alpha + \theta = 0$, then only partitions with a single component have non-zero probabilities:

$$P_N(\Lambda) = \begin{cases} 1, & \text{if } R_N = 1; \\ 0, & \text{otherwise.} \end{cases}$$

(1.8) If $n \equiv -\theta/\alpha = 1, 2, \dots$ and $\alpha \to -\infty$, $\theta \to \infty$ then the Ewens–Pitman distributions converge to the limiting distribution

$$P_N(\Lambda) = \frac{N!}{(1!)^{R_1} R_1!\,(2!)^{R_2} R_2! \dots}\,\frac{(n)_K}{n^N},$$

where $K = \sum R_j$ is the number of blocks in the partition Λ, and the notation $(n)_K = n(n-1)\dots(n-K+1)$ is used for factorial powers. It

[2] See also [20, 21, 24, 25]. – *A.G.*

is easy to see that

$$P_N(\Lambda) = \frac{N!}{(1!)^{R_1}\,(2!)^{R_2}\dots}\; m_\Lambda\Big(\overbrace{\frac{1}{n},\,\dots,\,\frac{1}{n}}^{n\text{ times}}\Big),$$

where m_Λ is the monomial symmetric function associated with the partition Λ (see [7] for basic notations and facts on symmetric functions).
(1.9) If $\alpha \to 1$ while the parameter $\theta > c > -1$ is bounded below (here a constant c does not depend on α), then only components of length 1 have positive probabilities in equation (1.3):

$$P_N(\Lambda) = \begin{cases} 1, & \text{if } \quad R_1 = N; \\ 0, & \text{otherwise.} \end{cases}$$

The same conclusion holds if $\theta \to \infty$ and $\alpha \geqslant 0$.
(1.10) Assume that $\alpha \to 1$, $\theta \to -1$ in such a way that the limits

$$p = \lim \frac{1-\alpha}{\theta+1}; \qquad q = \lim \frac{\theta+\alpha}{\theta+1}. \tag{1.11}$$

exist. Then the Ewens–Pitman distributions converge to the distribution

$$P_N(\Lambda) = \begin{cases} p, & \text{if } \quad R_N = 1; \\ q, & \text{if } \quad R_1 = N; \\ 0, & \text{otherwise.} \end{cases}$$

It is obvious that the conditions (1.11) are equivalent to the existence of the limiting distribution $P_2 = \lim P_2^{(\alpha,\theta)}$ when $\alpha \to 1$, $\theta \to -1$.

2. The urn model[3]. There is a natural urn model leading to the distribution (1.3). It will be convenient to replace "particles" by unit squares of different "colors". A fixed fraction α of some squares may be left blank. The color spectrum in a set of colored squares can be described by a partition $\Lambda = (R_1, R_2, \dots)$, in which R_j indicates the number of colors represented by exactly $j = 1, 2, \dots$ squares. At each stage of the process we add to the urn just one new square.

The process starts up at time $N = 1$ with a single square colored in, say, red, except for a fraction α which is painted white. Also, we put in the urn a nonstandard white rectangle with the area θ, see Fig. 2. (If

[3] The model describes a Markov chain now commonly known as the Dubins–Pitman "Chinese restaurant process" [25]. – *A.G.*

$\theta < 0$, we rather remove an appropriate white part of the initial square, so that the total white area equals $\theta + \alpha > 0$).

At each stage of the process we pick up a point at the surface of one of the squares currently in the urn. If the point is colored (i.e., not white), we add to the urn a new square all painted in exactly the same color as the point touched. If the point was white, we add a square which is partially white, and partially colored in an entirely new color, never represented in the urn before. The fraction of white is always assumed to be α.

(2.1) Lemma. *The formula* (1.3) *provides the probability to obtain the color spectrum Λ at the Nth stage of the urn model above.*

Proof. Assume that a state with the color spectrum $\Lambda \in \mathcal{Y}_N$ arose by adding a new colored square to j previously existing squares of this color. The probability of such a transition is[4].

$$p(\lambda, \Lambda) = \frac{(j-\alpha)\,(R_j+1)}{\theta + N - 1}. \tag{2.2}$$

On the other hand, the probability to obtain a square of a new color equals

$$p(\lambda, \Lambda) = \frac{\theta + K\alpha}{\theta + N - 1}, \tag{2.3}$$

where K denotes the number of different colors represented in the initial urn[5].

We use now the induction in N, assuming that the distribution of color spectrum in the urn is given by the Ewens–Pitman Formula. It is easy to check that

$$P_{N-1}^{(\alpha,\theta)}(\lambda)\; p(\lambda, \Lambda) = \frac{jR_j}{N}\; P_N^{(\alpha,\theta)}(\Lambda), \tag{2.4}$$

[4] (R_j) refers to Λ. – *A.G.*
[5] With $N-1$ squares. – *A.G.*

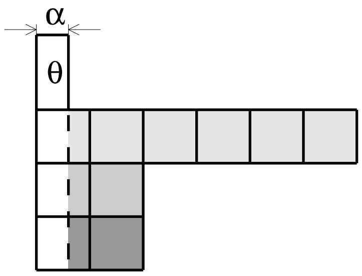

Fig. 2. An example urn contents after $N = 10$ moves.

provided that upon the transition from λ to Λ the new square becomes the jth one of its color.

It follows that the probability to obtain the spectrum Λ at the Nth stage equals

$$\sum_{\lambda \in \mathcal{Y}_{N-1}} P_{N-1}^{(\alpha,\theta)}(\lambda)\, p(\lambda, \Lambda) = P_N^{(\alpha,\theta)}(\Lambda) \tag{2.5}$$

and the Lemma is proved. □

3. The Ewens–Pitman distributions are coherent. Assume there is an urn containing N colored squares with the color spectrum $\Lambda \in \mathcal{Y}_N$. By erasing a square whose color is represented by exactly j squares in the urn, we obtain a new partition $\Lambda_{(j)} \in \mathcal{Y}_{N-1}$ with the color spectrum

$$N - 1 = r_1 + 2r_2 + 3r_3 + \dots, \tag{3.1}$$

where $r_{j-1} = R_{j-1} + 1$, $r_j = R_j - 1$ and $r_i = R_i$ for $i \neq j-1, j$. If all the squares in the urn have equal chances to be deleted, the probability to obtain the set $\Lambda_{(j)}$ with the color spectrum (3.1) equals

$$q(\Lambda_{(j)}, \Lambda) = \frac{jR_j}{N}. \tag{3.2}$$

The numbers (3.2) are called *cotransition probabilities*. One can use them to define, starting with a distribution P_N on the set $\mathcal{Y}_N$, a new distribution P_{N-1} on the set $\mathcal{Y}_{N-1}$ with probabilities

$$P_{N-1}(\lambda) = \sum_{\Lambda \in \mathcal{Y}_N} q(\lambda, \Lambda)\, P_N(\Lambda). \tag{3.3}$$

(3.4) Definition. *A family of probability distributions P_N on the sets of partitions $\mathcal{Y}_N$ is called coherent, if the identity (3.3) holds for all $N = 1, 2, \ldots$.*

The concept is identical with that of *partition structure*, introduced by Kingman in [13]. In a wider context of general branching diagrams the coherent families were systematically studied in [1, 12][6].

(3.5) Lemma. *For any fixed parameters $0 \leqslant \alpha < 1$, $\theta > -\alpha$, the Ewens–Pitman distributions given by the equation (1.3) form a coherent family.*

Proof. The Lemma follows directly from equation (2.4), since $\sum_\Lambda p(\lambda, \Lambda) = 1$. $\square$

4. The Kolchin model. We start with two integer valued random variables K, S with distributions

$$\mathrm{Prob}\{K = j\} = \varkappa_j; \quad j = 1, 2, \ldots \tag{4.1}$$

$$\mathrm{Prob}\{S = j\} = \sigma_j; \quad j = 1, 2, \ldots. \tag{4.2}$$

Consider a random vector $\mathbf{S} = (S_1, S_2, \ldots, S_K)$ such that its components are independent and identically distributed with the common distribution (4.2). We associate with this vector a random partition $\Lambda = (R_1, R_2, \ldots)$ of an integer $N = S_1 + S_2 + \ldots + S_K$, where R_j denotes the number of components of $\mathbf{S}$ equal to $j = 1, 2, \ldots$. Recall the identities

$$N = R_1 + 2R_2 + 3R_3 + \ldots; \qquad K = R_1 + R_2 + R_3 + \ldots$$

introduced earlier in Section 1.

By definition, the probability to obtain a partition Λ equals

$$P(\Lambda) = \varkappa_K \, \frac{K!}{R_1! R_2! \ldots} \, \sigma_1^{R_1} \sigma_2^{R_2} \ldots. \tag{4.3}$$

We are interested in conditional probabilities $P_N(\Lambda)$, for fixed value of N. According to (4.3),

$$P_N(\Lambda) = \frac{\varkappa_K}{C_N} \, \frac{K!}{R_1! R_2! \ldots} \, \sigma_1^{R_1} \sigma_2^{R_2} \ldots, \tag{4.4}$$

where

$$C_N = \mathrm{Prob}\{S_1 + \ldots + S_K = N\} = \sum_{\Lambda \in \mathcal{Y}_N} P(\Lambda)$$

is the normalizing constant.

[6] See also [21, 2, 22, 23, 11, 4]. – *A.G.*

(4.5) Definition. *We say that the distributions* (4.4) *are generated via Kolchin's model. The integer distributions* (4.1), (4.2) *are called the parameters of the model.*

5. Quasi-binomial distributions. As we shall show in the next Section 6, any Ewens–Pitman distribution can be generated via Kolchin's model. To this end, we shall presently introduce a specific class of integer distributions similar to binomial distributions. The parameter distributions in the Kolchin model leading to the formula (1.3) will be taken from this class.

(5.1) Definition. *We call an integer distribution* $\{p_1, p_2, \ldots\}$ *(where* $p_k = p_k(a, x)$, $k = 1, 2, \ldots$*) quasi-binomial, if the ratios*

$$\widetilde{p}_k = \frac{(k+1)p_{k+1}}{p_k}; \qquad k = 1, 2, \ldots \tag{5.2}$$

form an arithmetic series with parameters a, x*:*

$$\widetilde{p}_k = a + x\,k; \qquad k = 1, 2, \ldots. \tag{5.3}$$

In the next Section we use the following complete classification of quasi-binomial distributions. The space of admissible parameters is shown in Figure 3.

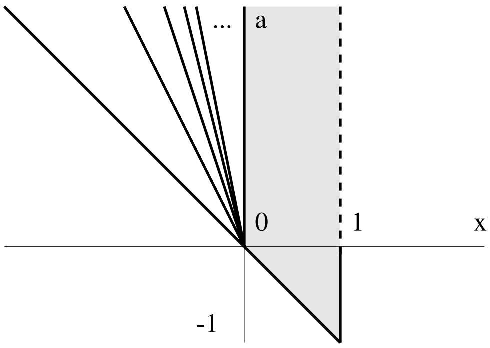

Fig. 3. The parameter space for quasi binomial distributions.

(5.4) Lemma. *Every quasi-binomial distribution can be obtained by one of the formulae (5.5)–(5.7):*

$$p_k(a, x) = \frac{1}{\left((1-x)^{-a/x} - 1\right)} \frac{a(a+x)\ldots(a+(k-1)x)}{k!}; \quad k = 1, 2, \ldots, \tag{5.5}$$

where $0 < x < 1,\ a > -x$ [7] *or* $x = 1,\ -1 < a < 0$*;*

$$p_k(a,0) = \frac{1}{\left(e^a - 1\right)} \frac{a^k}{k!}; \qquad k = 1, 2, \dots, \tag{5.6}$$

where $x = 0$ *and* $a > 0$*; finally*

$$p_k(-mx, x) = \frac{(-x)^k}{\left((1-x)^m - 1\right)} \binom{m}{k}; \quad k = 1, 2, \dots, m, \tag{5.7}$$

where $x < 0$ *and* $-a/x \equiv m = 1, 2, 3, \dots$*.*

Remark. For $a = 0$ the formula (5.5) is meaningless. In this case we replace (5.5) by the logarithmic distribution

$$p_k(0,x) = \frac{x^k}{-k\ln(1-x)}; \quad k = 1, 2, \dots \tag{5.8}$$

with the parameter $0 < x < 1$. So defined, the probabilities $p_k(a,x)$ depend on the parameters a, x continuously.

Proof. For any quasi-binomial distribution $p_k = p_k(a,x)$ there is an identity

$$k!\,\frac{p_k}{p_1} = \widetilde{p}_1\widetilde{p}_2\dots\widetilde{p}_{k-1} = \prod_{j=1}^{k-1}(a + jx).$$

Using the normalization condition $\sum_k p_k = 1$, we can write uniquely the probabilities p_k as functions of parameters a, x. The series $\sum_k p_k(a,x)$ converges in two cases only: if $|x| \leqslant 1$, or else if the series is finite, i.e., if $a + mx = 0$ for some $m = 1, 2, \dots$.

It is easy to see that the numbers p_k are only positive in the cases indicated in Lemma (5.4). In fact, we have

$$a + x = \widetilde{p}_1 = \frac{2p_2}{p_1} \geqslant 0; \qquad x = \lim_{k\to\infty} \frac{\widetilde{p}_k}{k} = \lim_{k\to\infty} \frac{(k+1)p_{k+1}}{kp_k} \leqslant 1,$$

since the probabilities p_k are positive and decrease at infinity. If $x < 0$, then $-a/x$ is an integer. Otherwise, we could choose an integer m, such that $m - 1 < -a/x < m$. But this would imply that $\widetilde{p}_m = a + mx < 0$ and $p_{m+1} < 0$. $\square$

[7] Or $a = -x$. – *A.G.*

In Section 6, we shall use the generating functions

$$\varphi^{(a,x)}(z) = \sum_{k=1}^{\infty} p_k(a,x)\, z^k$$

of quasi-binomial distributions. According to Lemma (5.4), they are given by equations

$$\varphi^{(a,x)}(z) = \frac{(1-xz)^{-a/x} - 1}{(1-x)^{-a/x} - 1}, \tag{5.9}$$

if $0 < x < 1$ and $a > -x$, $a \neq 0$;

$$\varphi^{(0,x)}(z) = \frac{\ln(1-xz)}{\ln(1-x)}, \tag{5.10}$$

if $0 < x < 1$ and $a = 0$;

$$\varphi^{(a,x)}(z) = \frac{e^{az} - 1}{e^{a} - 1}, \tag{5.11}$$

if $x = 0$, $a > 0$ and, finally,

$$\varphi^{(a,x)}(z) = \frac{(1-xz)^{m} - 1}{(1-x)^{m} - 1}, \tag{5.12}$$

if $x < 0$ and $a = -mx$, where $m = 1, 2, \ldots$.

6. All Ewens–Pitman distributions can be obtained via Kolchin's model. We show in this Section that, using appropriate quasi-binomial distributions (4.1), (4.2) as parameters in Kolchin's model, one can obtain all Ewens–Pitman distributions as conditional distributions (4.4). Let us start with a study of the random variable N.

(6.1) Lemma. *Assume that the random variables S, K in the Kolchin model have quasi-binomial distributions $\sigma_s = p_s(a,x)$, $\varkappa_k = p_k(b,y)$, and that $y = -a/\sigma_1$. Then the distribution of the variable* (1.1) *is quasi-binomial, too:*

$$C_n = \operatorname{Prob}\{N = n\} = p_n(-ab/y, x); \qquad n = 1, 2, \ldots. \tag{6.2}$$

Proof. One can see from (4.4) that the generating function $\varphi_N(z) = \sum_{n=1}^{\infty} C_n z^n$ of the integer distribution $\{C_1, C_2, \ldots\}$ is related to generating functions $\varphi_S(z)$, $\varphi_K(z)$ of distributions (4.1), (4.2) by the identity

$$\varphi_N(z) = \varphi_K(\varphi_S(z)).$$

We consider the three cases of Lemma (5.4) separately.

(i) If $0 < x < 1$ and $a > -x$,[8] then $y = -a/\sigma_1 = 1 - (1-x)^{-a/x}$,

[8] With $a \neq 0$. – *A.G.*

which implies $(1-y) = (1-x)^{-a/x}$. The generating functions for (4.1), (4.2) have the form

$$\varphi_S(z) = \frac{(1-xz)^{-a/x}-1}{(1-x)^{-a/x}-1}; \quad \varphi_K(z) = \frac{(1-yz)^{-b/y}-1}{(1-y)^{-b/y}-1}.$$

Since $(1-y\varphi_S(z)) = (1-xz)^{-a/x}$, we get for the generating function of N the expression

$$\varphi_N(z) = \varphi_K(\varphi_S(z)) = \frac{(1-xz)^{\frac{ab}{xy}}-1}{(1-x)^{\frac{ab}{xy}}-1},$$

corresponding to the quasi-binomial distribution $C_n = p_n(-ab/y, x)$. If $a = 0$, we have $\sigma_1 = -x/\ln(1-x)$, so that $y = -a/\sigma_1 = 0$ and the generating functions take the form

$$\varphi_S(z) = \frac{\ln(1-xz)}{\ln(1-x)}, \qquad \varphi_K(z) = \frac{e^{bz}-1}{e^b-1}.$$

Hence,

$$\varphi_N(z) = \varphi_K(\varphi_S(z)) = \frac{(1-xz)^{b/\ln(1-x)}-1}{(1-x)^{b/\ln(1-x)}-1},$$

and the variable N has the quasi-binomial distribution

$$C_n = p_n(-bx/\ln(1-x), x).$$

(ii) In the second case we assume that $x = 0$ and $a > 0$, so that $y = -a/\sigma_1 = 1-e^a < 0$ and $(1-y) = e^a$. Hence, $b = -my$, $m = 1, 2, \dots$ and the generating functions of the distributions (4.1), (4.2) take the form

$$\varphi_S(z) = \frac{e^{az}-1}{e^a-1}; \quad \varphi_K(z) = \frac{(1-yz)^m-1}{(1-y)^m-1}.$$

As a consequence, $(1-y\varphi_S(z)) = e^{-az}$ and we obtain for the generating function of the variable N the expression

$$\varphi_N(z) = \frac{e^{amz}-1}{e^{am}-1},$$

corresponding to the quasi-binomial distribution $C_n = p_n(am, 0)$.

(iii) In the last case of Lemma (5.4) we have $x < 0$ and $a = -kx$, $k = 1, 2, \dots$, which implies $y = -a/\sigma_1 = 1-(1-x)^k < 0$ and $(1-y) = (1-x)^k$. The generating functions have the form

$$\varphi_S(z) = \frac{(1-xz)^k-1}{(1-x)^k-1}; \quad \varphi_K(z) = \frac{(1-yz)^m-1}{(1-y)^m-1},$$

where $m = 1, 2, \ldots$. Hence, $(1 - y\varphi_S(z)) = (1 - xz)^k$ and

$$\varphi_N(z) = \frac{(1 - xz)^{km} - 1}{(1 - x)^{km} - 1}.$$

In this case the variable N has the quasi-binomial distribution $C_n = p_n(-kmx, x)$, too, and the lemma is proved. $\square$

We establish now a connection between quasi-binomial distributions and the Ewens–Pitman Formula.

(6.3) Proposition. *Let P_N denote the conditional distribution on the set of partitions $\mathcal{Y}_N$, obtained via Kolchin's model (i.e., using the equation (4.4)) with the parameter distributions*

$$\sigma_k = p_k(a, x), \quad \varkappa_k = p_k(b, y); \qquad k = 1, 2, 3, \ldots, \tag{6.4}$$

and assume that $y = -a/\sigma_1$. Then the distribution P_N is given by the Ewens–Pitman Formula (1.3) with the parameters

$$\alpha = -\frac{a}{x}; \qquad \theta = -\frac{ab}{xy}. \tag{6.5}$$

More precisely, there are the following particular cases:

(1) *if $0 < x < 1$, $-x < a < 0$ and $y = 1 - (1 - x)^{-a/x}$, $b > -y$, then $0 < \alpha < 1$ and $\theta > -\alpha$;*

(2) *if $0 < x < 1$, $a = 0$ and $y = 0$, $b > 0$, then $\alpha = 0$ and $\theta = -b/\ln(1 - x) > 0$;*

(3) *$0 < x < 1$, $a > 0$ and $y = 1 - (1 - x)^{-a/x}$, $m \equiv -b/y = 1, 2, \ldots$, then $\alpha < 0$ and $-\theta/\alpha = m = 1, 2, \ldots$;*

(4) *if $x = 0$, $a > 0$ and $y = 1 - e^a$, $m \equiv -b/y = 1, 2, \ldots$, then the distributions $P_N(\Lambda)$ coincide with the limiting case (1.8) of the equation (1.3).*

Proof. We consider the cases (1)–(4) separately.

(1) If $0 < x < 1$ and $-x < a < 0$, then $y = -a/\sigma_1 = 1-(1-x)^{-a/x} > 0$ and we can take, in the formula (6.4) for the second parametric distribution $\varkappa$, any value $b > -y$. Since $(1 - x)^{-a/x} = (1 - y)$, the normalizing constant in Lemma (6.1) equals

$$C_N = p_N(-ab/y, x) = \frac{1}{N!\big((1 - y)^{-b/y} - 1\big)} \prod_{j=0}^{N-1} (-ab/y + jx).$$

The equation (4.4) takes the form

$$P_N(\Lambda)=\frac{p_K(b,y)}{p_N(-ab/y,x)}\;\frac{K!}{R_1!R_2!\dots}\;\prod_{S\geqslant 1}p_S^{R_S}(a,x)=$$

$$=\frac{N!\,\big((1-y)^{-b/y}-1\big)}{\prod_{j=0}^{N-1}\big(-ab/y+jx\big)}\;\frac{\prod_{j=0}^{K-1}(b+jy)}{\big((1-y)^{-b/y}-1\big)\,K!}\times$$

$$\times\frac{K!\,x^N}{1^{R_1}R_1!\,2^{R_2}R_2!\dots}\left(\frac{a}{-xy}\right)^K\prod_{S\geqslant 1}\prod_{j=1}^{S-1}\left(1+\frac{a}{jx}\right)^{R_S}=$$

$$=d_\Lambda\;\frac{\prod_{i=0}^{K-1}(\theta+i\alpha)\;\prod_{j\geqslant 1}\prod_{i=1}^{j-1}\big(1-\alpha\,/\,i\big)^{R_j}}{\prod_{i=0}^{N-1}(\theta+i)},$$

where $\alpha=-a/x$ and $\theta=-ab\,/\,xy$.

(2) If $0<x<1$ and $a=0$, then $y=-a/\sigma_1=0$, and $b>0$ is any positive number. Using the notation $\theta=-b/\ln(1-x)$, we have $(1-x)^{-\theta}=e^b$ and

$$C_N=p_N(-bx/\ln(1-x),x)=\frac{\prod_{j=0}^{N-1}(x\theta+jx)}{N!\big(e^b-1\big)}.$$

The equation (4.4) implies

$$P_N(\Lambda)=\frac{p_K(b,0)}{p_N(\theta x,x)}\;\frac{K!}{R_1!R_2!\dots}\;\prod_{S\geqslant 1}p_S^{R_S}(0,x)=$$

$$=\frac{N!\,\big(e^b-1\big)}{\prod_{j=0}^{N-1}(\theta x+jx)}\;\frac{b^K}{\big(e^b-1\big)\,K!}\;\frac{K!\,x^N}{1^{R_1}R_1!\,2^{R_2}R_2!\dots}\;\frac{x^N\theta^K}{b^K}=\frac{d_\Lambda\,\theta^K}{[\theta]_N}$$

and the second case (2) is done.

(3) If $0<x<1$ and $a>0$, then $y=1-(1-x)^{-a/x}<0$. By Lemma (5.4), we can only choose the parameter b in such a way that the ratio $-b/y=m=1,2,\dots$ is integer. The remaining part of the proof is similar to that of item (1).

(4) Assume now that $x=0$ and $a>0$; then $y=1-e^a$ and once again we have to choose such b that $-b/y=m=1,2,\dots$. By Lemma (6.1),

$$C_N=p_N(am,0)=\frac{(am)^N}{N!\big(e^{am}-1\big)}.$$

Since $1-y=e^a$, we get

$$\begin{aligned} P_N(\Lambda) &= \frac{p_K(-my,y)}{p_N(am,0)}\,\frac{K!}{R_1!R_2!\dots}\prod_{S\geqslant 1} p_S^{R_S}(a,0) \\ &= \frac{N!\,(e^{am}-1)}{(am)^N}\,\frac{(-y)^K\,(m)_K}{(e^{am}-1)\,K!}\prod_{S\geqslant 1}\Big(\frac{a^S}{S!y}\Big)^{R_S} \\ &= \frac{N!}{(1!)^{R_1}R_1!\,(1!)^{R_2}R_2!\dots}\,\frac{(m)_K}{m^N}, \end{aligned}$$

which is identical with the equation (1.8). The proof of the proposition is completed. □

Remark. If one chooses the parameters of a quasi-binomial variable S to be $x<0$, $-a/x\equiv k=1,2,\dots$, then the only choice for the parameters of K would be $y=1-(1-x)^k<0$ and $-b/y\equiv m=1,2,\dots$. But it follows that $C_N=p_N(-kmx,x)$, and the variable N cannot exceed km. This case is of no interest to us.

7. The characterization of the Ewens–Pitman distributions. The family of distributions (1.3) is coherent (see Section 3), and can be obtained via Kolchin's model (Section 4). We shall show now that these two properties characterize the Ewens–Pitman Formula uniquely.

(7.1) Theorem. *Consider the distribution P_N on the set of partitions $\mathcal{Y}_N$, $N=1,2,\dots$, obtained via Kolchin's model, and assume that the family $\{P_N\}$ is coherent. Then the distributions $P_N=P_N^{(\alpha,\theta)}$ are given by the Ewens–Pitman formula* (1.3). *There are two possibilities for the parameters α,θ:*

$$\text{either}\quad 0\leqslant\alpha<1\quad\text{and}\quad\theta>-\alpha\,; \tag{7.2}$$

$$\text{or}\quad \alpha=-\frac{\theta}{m},\quad\text{with}\quad m=1,2,\dots\quad\text{and}\quad\theta>0. \tag{7.3}$$

Also, a limiting case $\theta\to\infty$ can occur[9], *so that*

$$P_N(\Lambda)=\begin{cases}1, & \text{if } R_1=N;\\ 0, & \text{otherwise.}\end{cases} \tag{7.4}$$

[9] And the limiting cases of "coupon-collectors" partitions (1.8) which correspond to case (4) of Proposition 6.3. – *A.G.*

Proof. We fix a partition $\Lambda = (R_1, R_2, \dots) \in \mathcal{Y}_N$ and denote by $\Lambda^{(j)} \in \mathcal{Y}_{N+1}$ the partition with a term of Λ of length j increased by one. Thus,

$$\Lambda^{(j)} = (R_1, \dots, R_{j-1}, R_j - 1, R_{j+1} + 1, R_{j+2}, \dots),$$

and we also put $\Lambda^{(0)} = (R_1 + 1, R_2, R_3 \dots)$. Of course, for $j \geqslant 1$ the partition $\Lambda^{(j)}$ makes sense if $R_j > 1$[10] only. The partition $\Lambda^{(j)}$ has the same number of parts K as Λ, if $j \geqslant 1$, and one extra part, if $j = 0$. The coherency condition means that

$$\frac{R_1 + 1}{N + 1} P_{N+1}(\Lambda^{(0)}) + \sum_{j \geqslant 1} \frac{(j+1)(R_{j+1} + 1)}{N + 1} P_{N+1}(\Lambda^{(j)}) = P_N(\Lambda) \quad (7.5)$$

for all partitions $\Lambda \in \mathcal{Y}_N$.

(7.6) Lemma. *The coherency condition for the distributions* (4.4) *obtained via the Kolchin model is equivalent to the identity*

$$\frac{(K+1)\varkappa_{K+1}}{\varkappa_K} \sigma_1 + \sum_{j \geqslant 1} R_j \frac{(j+1)\sigma_{j+1}}{\sigma_j} = \frac{(N+1) C_{N+1}}{C_N}, \qquad (7.7)$$

where $\Lambda = (R_1, R_2, \dots) \in \mathcal{Y}_N$ *is any partition, and the numbers* N, K *are determined by multiplicities* $R_1, R_2, \dots$ *by the equations* (1.1), (1.2).

Proof. It follows immediately from (4.4), that

$$\frac{C_{N+1}}{C_N}(R_1 + 1) P_{N+1}(\Lambda^{(0)}) = \frac{(K+1)\varkappa_{K+1}}{\varkappa_K} \sigma_1 \, P_N(\Lambda)$$

and

$$\frac{C_{N+1}}{C_N}(j+1)(R_{j+1} + 1) P_{N+1}(\Lambda^{(j)}) = R_j \frac{(j+1)\sigma_{j+1}}{\sigma_j} \, P_N(\Lambda)$$

for admissible values of $j \geqslant 1$. If we substitute these identities in equation (7.5) and cancel out the common factor $P_N(\Lambda)$, the condition (7.7) follows. □

(7.8) Lemma. *Assume that* (7.7) *holds for all partitions* $\Lambda = (R_1, R_2, \dots)$. *Then the numbers*

$$\widetilde{\sigma}_j = \frac{(j+1)\sigma_{j+1}}{\sigma_j}; \qquad j = 1, 2, \dots \qquad (7.9)$$

[10] Correction: $R_j \geq 1$. – *A.G.*

form an arithmetic sequence:

$$\widetilde{\sigma}_j = x\,j + a; \qquad j = 1, 2, \dots \tag{7.10}$$

for appropriate values of x and a.

Proof. Assume that for some $i \geqslant 2$ the numbers R_{i-1}, R_{i+1} are positive. Consider a new partition $\mathrm{M} \in \mathcal{Y}_N$ with multiplicities R_{i-1}, R_i, R_{i+1} replaced by $R_{i-1} - 1, R_i + 2, R_{i+1} - 1$ correspondingly (all other multiplicities remain unchanged). In terms of Young diagrams, the transition from Λ to M simply means that a box is deleted from $(i+1)$th column of Λ and attached to its ith column. Note that the partitions Λ and M have equal numbers of components K. We see, by subtracting from the identity (7.7) a similar identity for the diagram M, that

$$\widetilde{\sigma}_{i+1} - \widetilde{\sigma}_i = \widetilde{\sigma}_i - \widetilde{\sigma}_{i-1}; \qquad i \geqslant 2$$

and Lemma (7.8) follows. $\square$

(7.11) Lemma. *It follows from the equation* (7.7) *that the ratios*

$$\widetilde{\varkappa}_K = \frac{(K+1)\varkappa_{K+1}}{\varkappa_K}; \qquad K = 1, 2, \dots \tag{7.12}$$

form an arithmetic sequence:

$$\widetilde{\varkappa}_K = y\,K + b, \qquad K = 1, 2, \dots; \tag{7.13}$$

moreover, the coefficients in (7.10), (7.13) *satisfy the identity* $a + y\sigma_1 = 0$.

Proof. Denote by

$$\mathrm{M} = (R_1 + 1, R_2, \dots, R_{i-2}, R_{i-1} + 1, R_i - 1, R_{i+1}, \dots) \in \mathcal{Y}_N$$

the partition obtained from Λ by replacing the multiplicities R_1, R_{i-1}, R_i by $R_1 + 1, R_{i-1} + 1, R_i - 1$ correspondingly (i.e., one box is removed from the ith column of the Young diagram Λ, and attached to its first column). It is clear that the partition M has one more component than Λ. Comparing the identities (7.7) for the partitions Λ and M we conclude that

$$(\widetilde{\varkappa}_{K+1} - \widetilde{\varkappa}_K)\sigma_1 = \widetilde{\sigma}_i - \widetilde{\sigma}_{i-1} - \sigma_1 = -a.$$

Hence, $\widetilde{\varkappa}_K = -Ka/\sigma_1 + b$ for appropriate b, and the Lemma (7.11) is proved. $\square$

The proof of Theorem 7.1 follows immediately from Proposition 6.3 and the lemmas of this section. $\square$

References

1. A. M. Vershik and S. V. Kerov, *Locally semisimple algebras. Combinatorial theory and the K_0-functor.* — J. Sov. Math. **38** (1987), 1701–1733.
2. S. V. Kerov, *Subordinators, and permutations acting with quasi-invariant measure* — Zap. Nauchn. Semin. POMI **23** (1995), 181–218.
3. S. V. Kerov, *Small cycles of big permutations.* — PDMI Preprint **2** (1995), 1–10.
4. S. V. Kerov and N. V. Tsilevich, *Stick breaking process generates virtual permutations with Ewens distribution.* — Zap. Nauchn. Semin. POMI **23** (1995), 162–180.
5. V. F. Kolchin, *A problem on allocations of particles into boxes, and the cycles of random permutations.* — Teor. Veroyatn. Primen. **16** (1971), 67–82.
6. V. F. Kolchin, B. A. Sevast'yanov, and V. P. Chistyakov, *Random Allocations.* Wiley, New York (1978).
7. I. G. Macdonald, *Symmetric Functions and Hall Polynomials.* Clarendon Press, Oxford (1995).
8. W. J. Ewens, *The sampling theory of selectively neutral allels.* — Theoret. Population Biol. **3** (1972), 87–112.
9. W. J. Ewens, *Population genetics theory – the past and the future.* — In: S. Lessard (ed.), Mathematical and Statistical Developments of Evolutionary Theory. Kluwer, Dordrecht (1990), pp. 177–27.
10. P. R. Halmos, *Random alms.* — Ann. Math. Stat. **15** (1944), 182–189.
11. S. Kerov, G. Olshanski, and A. Vershik, *Harmonic analysis on the infinite symmetric group.* — C. R. Acad. Sci. Paris **316** (1993), 773–778.
12. S. V. Kerov and A. M. Vershik, *The Grothendieck group of the infinite symmetric group and symmetric functions with the elements of the K_0-functor theory of AF-algebras.* — In: Representation of Lie Groups and Related Topics. Adv. Stud. Contemp. Math. **7**, Gordon and Breach, 1990, pp. 36–114.
13. J. F. C. Kingman, *The representation of partition structures.* — J. London Math. Soc. (2) **18** (1978), 374–380.
14. J. F. C. Kingman, *Random partitions in population genetics.* — Proc. R. Soc. Lond. A **361** (1978), 1–20.
15. M. Perman, J. Pitman, and M. Yor, *Size-biased sampling of Poisson point processes and excursions.* — Probab. Theory Related Fields **92** (1992), 21–39.
16. J. Pitman, *The two-parameter generalization of Ewens' random partition structure.* — Preprint, Department of Statistics, University of California at Berkeley **345** (1992), 1–23.
17. J. Pitman and M. Yor, *Arcsine laws and interval partitions derived from a stable subordinator.* — Proc. London Math. Soc. (3) **65** (1992), 326–356.
18. J. Pitman and M. Yor, *The two-parameter Poisson–Dirichlet distribution derived from a stable subordinator.* — Ann. Probab. **25** (1997), 855–900.
19. S. Tavaré and W. J. Ewens, *The Ewens Sampling Formula.* — In: Multivariate Discrete Distributions, N. L. Johnson, S. Kotz, and N. Balakrishnan (eds.), Wiley, New York, 1995, pp. 1–20.
20. R. A. Arratia, A. D. Barbour, and S. Tavaré, *Logarithmic Combinatorial structures: A Probabilistic Approach.* EMS (2003).
21. A. Gnedin and J. Pitman, *Exchangeable Gibbs partitions and Stirling triangles.* — This issue (2005).

22. S. V. Kerov, *Asymptotic Representation Theory of the Symmetric Group and Its Applications.* Amer. Math. Soc., Providence, RI, 2003.
23. S. V. Kerov, A. Okounkov, and G. Olshanski, *The boundary of Young graph with Jack edge multiplicities.* — Int. Math. Res. Not. (1998), 173–199.
24. J. Pitman, *Exchangeable and partially exchangeable random partitions.* — Probab. Theory Related Fields **102** (1995), 145–158.
25. J. Pitman, *Combinatorial Stochastic Processes* (Lecture Notes for St. Flour Course, July 2002). Springer Lect. Notes Math. (to appear).