The two-parameter generalization of Ewens' random partition structure

Jim Pitman *
Technical Report No. 345
Department of Statistics
U.C. Berkeley CA 94720
March 25, 1992
Reprinted with an appendix and updated references

July 29, 2003

1 Introduction

The sampling formula of Ewens [8] defines a probability distribution over unordered partitions of a positive integer n as follows: for a sequence of non-negative integers $(m_1, \ldots, m_n)$ with $\sum_i im_i = n$, the probability that the partition has m_i parts of size i, $i = 1, \ldots, n$, is

$$P_{n,\theta}(m_1, \ldots, m_n) = \frac{\theta^k}{\theta^{(1;n)}} \frac{n!}{\prod_{j=1}^n j^{m_j} m_j!} \tag{1}$$

where $k = m_1 + \ldots + m_n$ is the total number of parts, $\theta > 0$ is a parameter of the distribution, and $\theta^{(1;n)} = \theta(\theta + 1)\ldots(\theta + n - 1)$. For background and applications to genetics see [11, 12, 13, 6, 5, 19]. This note presents a family of distributions $P_{n,\theta,\alpha}$ for random partitions of n, with two parameters $\theta \geq 0$ and $0 \leq \alpha \leq 1$. For $\alpha = 0$, $\theta > 0$, $P_{n,\theta,0} = P_{n,\theta}$ as in (1). As indicated at the end of this introduction, the case $0 < \alpha < 1$ arises naturally in the study of certain random partitions associated with stable processes with index α. In particular the case $\alpha = 1/2$ is related to random partitions induced by the zeros of Brownian motion.

*Research supported by N.S.F. Grant MCS91-07531

Here is the formula for the partition distribution $P_{n,\theta,\alpha}$: for non-negative integers $(m_1, \ldots, m_n)$ with $\sum_i im_i = n$,

$$P_{n,\theta,\alpha}(m_1, \ldots, m_n) = N(m_1, \ldots, m_n) \frac{\theta^{(\alpha;k-1)}}{(\theta + \bar{\alpha})^{(1;n-1)}} \prod_{j=1}^{n} [\bar{\alpha}^{(1;j-1)}]^{m_j} \tag{2}$$

where $\bar{\alpha} = 1 - \alpha$;

$$N(m_1, \ldots, m_n) = \frac{n!}{[j!]^{m_j}\ m_j!}$$

is the number of partitions of a set of n elements into m_i classes of size $i, i = 1, 2, \ldots, n$; and for a real number a and non-negative integer m,

$$\theta^{(a;m)} = \begin{cases} 1 \text{ for } m = 0 \\ \theta(\theta + a) \ldots (\theta + (m-1)a) \text{ for } m = 1, 2, \ldots \end{cases}$$

Proposition 1 *For each $\theta > 0$ and $0 \leq \alpha < 1$, formula (2) defines a probability distribution on unordered partitions of n. These distributions are consistent in the sense of Kingman [17]: if a set of $n+1$ elements is partitioned into random subsets with sizes distributed according to $P_{n+1,\theta,\alpha}$, and independently one of the $n+1$ elements picked uniformly at random is deleted, the induced partition of n elements is distributed according to $P_{n,\theta,\alpha}$.*

There are two trivial cases of Proposition 1. For $\theta = 0$ the distribution $P_{n,0,\alpha}$ is concentrated on the partition with a single component of size n. For $\alpha = 1$ the distribution $P_{n,\theta,1}$ is concentrated on the partition with n components of size 1. For $\theta > 0$, $0 \leq \alpha < 1$, formula (2) assigns strictly positive probability to all possible partitions of n. In this case Proposition 1 is an immediate consequence of the next proposition:

Proposition 2 *Fix $\theta > 0$, $0 \leq \alpha < 1$. For $i = 1, 2, \ldots$ let X_i be independent random variables with beta $(\bar{\alpha}, \theta - \alpha + i\alpha)$ distributions. Let (P_i) be the random discrete probability distribution on $\{1, 2, \ldots\}$ defined as follows:*

$$P_i = \bar{X}_1 \bar{X}_2 \ldots \bar{X}_{i-1} X_i \tag{3}$$

where $\bar{x} = 1 - x$. Given (P_i), let Y_1, $Y_2, \ldots$ be independent and identically distributed on $\{1, 2, \ldots\}$ according to (P_i). Define an equivalence relation $\stackrel{n}{\sim}$ on $\{1, \ldots n\}$ by $i \stackrel{n}{\sim} j$ iff $Y_i = Y_j$. Then the random partition of n induced by $\stackrel{n}{\sim}$ has distribution $P_{n,\theta,\alpha}$ as in (2).

Proposition 2 effectively determines Kingman's representation [18] of the partition structure defined by formula (2). Proposition 2 is derived in Section 2 as a consequence of the following result.

Proposition 3 *For (P_i) and (Y_n) as in Proposition 2, let*

$$\begin{aligned} T_1 &= 1 \\ T_{n+1} &= \inf\{m > T_n : Y_m \notin \{Y_{T_1}, \ldots, Y_{T_n}\}, \end{aligned}$$

the sequence of indices at which new Y-values appear. Then

$$(P_{T_1}, P_{T_2}, \ldots) \stackrel{d}{=} (P_1, P_2, \ldots).$$

In other words, the random probability distribution defined by (3) is invariant under size biased random permutation. Proposition 3 is an immediate consequence of the work of Perman, Pitman and Yor [20, 21], who showed that a sequence of random variables with representation (3) is obtained by size biased sampling of a certain point process derived from a Poisson point process on $(0, \infty)$. Another derivation of Proposition 3 is given in Pitman[23].

Let $B = (B_t, 0 \leq t \leq 1)$ be a stochastic process, for example a Brownian motion or Brownian bridge. Independent of B let $U_1, \ldots, U_n$ be i.i.d. uniform $[0,1]$. Define an equivalence relation $\stackrel{n}{\sim}$ on $\{1, \ldots, n\}$ by $i \stackrel{n}{\sim} j$ iff U_i and U_j fall in the same excursion of B away from zero: that is to say $B_t \neq 0$ for all t between U_i and U_j. Let P_n be the distribution of the unordered partition of n induced by $\stackrel{n}{\sim}$. According to the results of [21], if B is any of the processes considered below, then the size-biased presentation of the lengths of maximal subintervals of $[0,1]$ that are free of zeros of B defines a random discrete distribution (P_i) of the form described in Proposition 2. Consequently, Proposition 2 implies

- If B is standard Brownian motion, then $P_n = P_{n,1/2,1/2}$
- If B is standard Brownian bridge, then $P_n = P_{n,1,1/2}$

Somewhat more generally,

- If B is a Bessel process of dimension δ, $0 < \delta < 2$, started at $B_0 = 0$, then $P_n = P_{n,\alpha,\alpha}$ for $\alpha = 1 - \delta/2$
- If B is a Bessel bridge of dimension δ, $0 < \delta < 2$, starting and ending at 0, then $P_n = P_{n,2\alpha,\alpha}$ for $\alpha = 1 - \delta/2$

Corollary 3.15 of [21] shows how to construct for any $0 < \alpha < 1$ and $\theta > 0$ a process B absolutely continuous with respect to a Bessel bridge of dimension $\delta = 1 - 2\alpha$ such that the zeros of B induce the partition distribution $P_{n,\theta,\alpha}$.

2 Derivation of the Formula

Let (P_i) be as in (3), and (Y_n) an iid sample from (P_i) as in the statement of Proposition 2. Formula (1) for the distribution of unordered partition of n induced by the values of $Y_1, \ldots, Y_n$ is a consequence of the following formula for the distribution of the ordered partition induced by the same values, with order defined by the order in which values appear in the sequence. The following proposition extends a formula due to Donnelly and Tavare [7] in case $\alpha = 0$.

Proposition 4 *Fix n, θ, α, and for $(n_1, \ldots, n_k)$ a sequence of integers with $1 \leq n_i \leq n$, $\sum_{i=1}^k n_i = n$, let $Q_{n,\alpha}(n_1, \ldots, n_k)$ denote the probability that for each $1 \leq j \leq n$, the jth value to appear in the Y sequence appears n_j times among $Y_1, \ldots, Y_n$. Then*

$$Q_{n,\theta,\alpha}(n_1, \ldots, n_k) = \#(n_1, \ldots, n_k) \frac{\theta^{(\alpha;k-1)} \prod_{i=1}^k \bar{\alpha}^{(1;n_i-1)}}{(\bar{\alpha}+\theta)^{(1;n-1)}} \tag{4}$$

where

$$\#(n_1, \ldots, n_k) = \frac{n!}{n_k(n_k+n_{k-1}) \ldots (n_k + \ldots + n_1) \ \prod_{i=1}^k (n_i - 1)!} \tag{5}$$

Proof. It is elementary that $\#(n_1, \ldots, n_k)$ as in (4) is the number of different ways to arrange n_1 values of one type, n_2 of a second, and so on, subject to the constraint that the first value is of the first type, the next distinct value of the second type, and so on. The other factor in (4) is the probability of any given arrangement of this kind. Indeed, for any given arrangement, due to the conclusion of Proposition 3, by conditioning on the successive P-values to appear this probability is found to be

$$\begin{aligned} &m_1(n_1 - 1, n_2 + \ldots + n_k) \ m_2(n_2 - 1, n_3 + \ldots + n_k) \ \ldots \\ &\ldots \ m_{k-1}(n_{k-1} - 1, n_k) \ m_k(n_k - 1, 0) \end{aligned} \tag{6}$$

where

$$\begin{aligned} m_i(r,s) &= E(X_i^r \bar{X}_i^s) && (7) \\ &= \frac{B(\bar{\alpha}+r, \theta - \alpha + i\alpha + s)}{B(\bar{\alpha}, \theta - \alpha + i\alpha)} && (8) \end{aligned}$$

with

$$B(a,s) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \tag{9}$$

the beta function. To illustrate the derivation of (6) for $k = 3$, the probability of any particular arrangement of kind $(n_1, \ldots, n_k)$ is found to be

$$E[X_1^{n_1-1}\bar{X}_1(\bar{X}_1X_2)^{n_2-1}(\bar{X}_1\bar{X}_2)(\bar{X}_1\bar{X}_2X_3)^{n_3-1}]$$

$$= E[X_1^{n_1-1}\bar{X}_1^{1+n_2-1+1+n_3-1}X_2^{n_2-1}\bar{X}_2^{1+n_3-1}X_3^{n_3-1}]$$
$$= E[X_1^{n_1-1}\bar{X}_1^{n_2+n_3}]E[X_2^{n_2-1}\bar{X}_2^{n_3}]E[X_3^{n_3-1}].$$

A similar calculation yields (6) for any other value of k. Now substituting (8) in (6) and using (9) and $\Gamma(r+1) = r\Gamma(r)$ allows (6) to be manipulated into the form shown in (4).

Proof of Proposition 2. what must be shown is that when formula (4) is summed over the $k!/(m_1!m_2!\ldots m_n!)$ distinct $(n_1, \ldots, n_k)$ such that $\#\{i : n_i = j\} = m_j$, the result is formula (1). In case $\alpha = 0$ this is a well known fact which amounts to the identity

$$\sum \frac{1}{n_k}\frac{1}{(n_k+n_{k-1})}\cdots\frac{1}{(n_k+\ldots n_1)} = \prod_{i=1}^{n}\frac{1}{i^{m_i}m_i!}, \tag{10}$$

where the sum is over the same range. (This identity is easily understood by computing the probability that a random permutation of n elements has m_i cycles of length i in two different ways).

The result in case $0 < \alpha < 1$ is obtained by using the identity (10) to simplify the sum in that case, after noting that for every $(n_1, \ldots n_k)$ whose distribution is given by $(m_1, \ldots, m_n)$,

$$\prod_{i=1}^{k}(n_i-1)! = \prod_{j=1}^{n}[(j-1)!]^{m_j}$$

and similarly

$$\prod_{i=1}^{k}\bar{\alpha}^{(1;n_i-1)} = \prod_{j=1}^{n}[\bar{\alpha}^{(1;j-1)}]^{m_j}$$

so both these products are constants so far as the summation is concerned.

3 Questions

According to Kingman [17], the Ewens sampling distributions $(P_{n,\theta,0})$ are characterized among all consistent families (P_n) by the following feature:

For a random partition with distribution P_n, if the part containing a point picked uniformly from $\{1,\ldots,n\}$, independently of the random partition is deleted, then given that k points remain, the distribution of the remaining partition of k is P_k, for every $1 \leq k \leq n$.

A corresponding property of $(P_{n,\theta,\alpha})$, follows easily from Proposition 2 and 3.

For a random partition with distribution $P_{n,\theta,\alpha}$, if the part containing a point picked uniformly from $\{1,\ldots,n\}$, independently of the random partition, is deleted, then given that k points remain, the distribution of the remaining partition of k is $P_{k,\theta+\alpha,\alpha}$, for every $1 \leq k \leq n$.

It would be interesting if this property could be used to characterize the family $(P_{n,\theta,\alpha})$. This motivates the following questions:

Question 5 *Suppose that (P_n) and (Q_n) are two consistent partition distributions, and that for every $n \geq 2$, when the part containing a random element is deleted from a P_n partition, given that k elements remain these are partitioned according to Q_k, $1 \leq k \leq n$. Does this imply $P_n = P_{n,\theta,\alpha}$ for some $\theta \geq 0$, $0 \leq \alpha < 1$?*

Question 6 *Suppose for each $i = 0, 1, 2, \ldots$ that (P_n^i) is a consistent family of partition distributions, and that for every $i \geq 1$ and $n \geq 2$, when the part containing a random element is deleted from a P_n^i partition, given that k elements remain these are partitioned according to P_k^{i+1}, $1 \leq k \leq n$. Does this imply $P_n^0 = P_{n,\theta,\alpha}$ (hence $P_n^i = P_{n,\theta+i\alpha,\alpha}$) for some $\theta \geq 0$, $0 \leq \alpha \leq 1$?*

By arguing as in Hoppe [12], using Kingman's representation of the partition distributions, these questions have the same answers as the next two questions respectively:

Question 7 *Does the two parameter family of joint distributions for (X_i) described in Proposition 2 comprise all joint distributions for a sequence of random variables (X_i) with values in $[0,1]$ such that both X_1 is independent of $(X_2, X_3, \ldots)$, and (P_i) defined by (3) is invariant under size biased permutation?*

Question 8 *Are the beta distributions for X_i displayed in Proposition 2 the only possible distributions for independent X_i such that (P_i) defined by (3) is invariant under size biased permutation?*

It appears possible to construct an example to show that the answer to Question 7 is no, hence also the answer to Question 5 is no. Some results related to Question 8 may be found in Pitman [23].

A Subsequent literature

The results of this technical report were published in [22]. The two-parameter family was characterized in various ways in [16, 23, 32]. Applications to Brownian and Bessel excursions appear in [25]. The associated two-parameter family of Poisson-Dirichlet distributions was decribed in [29]. The family has found applications in the theory of processes of fragmentation and coagulation [1, 2, 3, 26]. Other papers describing various aspects of the two-parameter family are [9, 30, 31]. Various generalizations are treated in [24], [28], [10]. Much of this literature is reviewed in [27]. See Carlton [4], [14], [15] for related studies from the perspective of Bayesian non-parametric statistics.

References

[1] D.J. Aldous and J. Pitman. The standard additive coalescent. *Ann. Probab.*, 26:1703–1726, 1998.

[2] J. Bertoin and J. Pitman. Two coalescents derived from the ranges of stable subordinators. *Electron. J. Probab.*, 5:no. 7, 17 pp., 2000.

[3] E. Bolthausen and A.-S. Sznitman. On Ruelle's probability cascades and an abstract cavity method. *Comm. Math. Phys.*, 197(2):247–276, 1998.

[4] M. A. Carlton. *Applications of the two-paramter Poisson-Dirichlet distribution*. PhD thesis, U.C.L.A., 1999.

[5] R. A. Doney. On the asymptotic behaviour of first passage times for transient random walk. *Probability Theory and Related Fields*, 81:239 – 246, 1989.

[6] P. Donnelly. Partition structures, Pólya urns, the Ewens sampling formula, and the ages of alleles. *Theoretical Population Biology*, 30:271 – 288, 1986.

[7] P. Donnelly and S. Tavaré. The ages of alleles and a coalescent. *Adv. Appl. Probab.*, 18:1–19 & 1023, 1986.

[8] W.J. Ewens. The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.*, 3:87 – 112, 1972.

[9] S. Feng and F. M. Hoppe. Large deviation principles for some random combinatorial structures in population genetics and Brownian motion. *Ann. Appl. Probab.*, 8(4):975–994, 1998.

[10] A. Gnedin and J. Pitman. Regenerative composition structures. Technical Report 644, Dept. Statistics, U.C. Berkeley, 2003. Available at http://www.stat.berkeley.edu/tech-reports/.

[11] F. M. Hoppe. Pólya-like urns and the Ewens sampling formula. *Journal of Mathematical Biology*, 20:91 – 94, 1984.

[12] F. M. Hoppe. Size-biased filtering of Poisson-Dirichlet samples with an application to partition structures in genetics. *Journal of Applied Probability*, 23:1008 – 1012, 1986.

[13] F. M. Hoppe. The sampling theory of neutral alleles and an urn model in population genetics. *Journal of Mathematical Biology*, 25:123 – 159, 1987.

[14] L. F. James. Generalized weighted gamma random measures and the two-parameter Poisson Dirchlet distribution. Preprint, 2001.

[15] L. F. James. Poisson calculus for spatial neutral to the right processes. arXiv:math.PR/0305053, 2003.

[16] S. Kerov. Coherent random allocations and the Ewens-Pitman formula. PDMI Preprint, Steklov Math. Institute, St. Petersburg, 1995.

[17] J. F. C. Kingman. Random partitions in population genetics. *Proc. R. Soc. Lond. A.*, 361:1–20, 1978.

[18] J. F. C. Kingman. The representation of partition structures. *J. London Math. Soc.*, 18:374–380, 1978.

[19] J. F. C. Kingman. *The Mathematics of Genetic Diversity.* SIAM, 1980.

[20] M. Perman. Order statistics for jumps of normalized subordinators. *Stoch. Proc. Appl.*, 46:267–281, 1993.

[21] M. Perman, J. Pitman, and M. Yor. Size-biased sampling of Poisson point processes and excursions. *Probab. Th. Rel. Fields*, 92:21–39, 1992.

[22] J. Pitman. Exchangeable and partially exchangeable random partitions. *Probab. Th. Rel. Fields*, 102:145–158, 1995.

[23] J. Pitman. Random discrete distributions invariant under size-biased permutation. *Adv. Appl. Prob.*, 28:525–539, 1996.

[24] J. Pitman. Some developments of the Blackwell-MacQueen urn scheme. In T.S. Ferguson et al., editor, *Statistics, Probability and Game Theory; Papers in honor of David Blackwell*, volume 30 of *Lecture Notes-Monograph Series*, pages 245–267. Institute of Mathematical Statistics, Hayward, California, 1996.

[25] J. Pitman. Partition structures derived from Brownian motion and stable subordinators. *Bernoulli*, 3:79–96, 1997.

[26] J. Pitman. Coalescents with multiple collisions. *Ann. Probab.*, 27:1870–1902, 1999.

[27] J. Pitman. Combinatorial stochastic processes. Technical Report 621, Dept. Statistics, U.C. Berkeley, 2002. Lecture notes for St. Flour course, July 2002. Available via www.stat.berkeley.edu.

[28] J. Pitman. Poisson-Kingman partitions. In D. R. Goldstein, editor, *Science and Statistics: A Festschrift for Terry Speed*, volume 30 of *Lecture Notes-Monograph Series*, pages 1–34. Institute of Mathematical Statistics, Hayward, California, 2003.

[29] J. Pitman and M. Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann. Probab.*, 25:855–900, 1997.

[30] H. Yamato and M. Sibuya. Moments of some statistics of Pitman sampling formula. *Bull. Inform. Cybernet.*, 32(1):1–10, 2000.

[31] H. Yamato, M. Sibuya, and T Nomachi. Ordered sample from two-parameter gem distribution. *Statist. Probab. Lett.*, 55:19–27, 2001.

[32] S.L. Zabell. The continuum of inductive methods revisited. In J. Earman and J. D. Norton, editors, *The Cosmos of Science*, Pittsburgh-Konstanz Series in the Philosophy and History of Science, pages 351–385. University of Pittsburgh Press/Universitätsverlag Konstanz, 1997.