The Ewens Sampling Formula

Simon Tavaré
University of Southern California *

W.J. Ewens
University of Pennsylvania

April 12, 1995

To appear as *'The Ewens Multivariate Distribution'* in
Multivariate Discrete Distributions
by N.L. Johnson, S. Kotz and N. Balakrishnan (Wiley).

*ST: Departments of Mathematics and Biological Sciences, University of Southern California, Los Angeles, CA 90089-1113. ST was supported in part by NSF grant DMS90-05833. WJE: Department of Biology, University of Pennsylvania, Philadelphia, PA 19104, and Department of Mathematics, Monash University, Clayton, Vic 3168, Australia. WJE was supported in part by NIH grant GM21135. The authors would like to thank Jim Pitman for his help. His comments on an earlier draft of this chapter led to substantial improvements in the presentation.

1 Genesis and History

The Ewens Multivariate Distribution, called in genetics the Ewens Sampling Formula (ESF), describes a specific probability for the partition of the positive integer n into parts. It was discovered by Ewens (1972) as providing the probability of the partition of a sample of n selectively equivalent genes into a number of different gene types (alleles), either exactly in some models of genetic evolution or as a limiting distribution (as the population size becomes indefinitely large) in others. It was also derived by Antoniak (1974) in the context of Bayesian statistics.

The impetus for the derivation of the formula came from the non-Darwinian theory of evolution. It is claimed, under this theory, that the quite extensive genetical variation observed in natural populations is, on the whole, not due to natural selection, but arises rather as a result of purely stochastic changes in gene frequency in finite populations. The ESF describes the partition distribution of a sample of n genes into allelic types when there are no selective differences between types, and thus provides the null hypothesis distribution for the non-Darwinian theory.

The distribution contains one parameter, usually denoted by θ, which in the genetic context is related to the rate of mutation of the genes to new allelic types, to the population size and to the details of the evolutionary model, being extremely robust with respect to these details. For the case $\theta = 1$ the distribution is quite old, going back in effect to Cauchy, since it then describes the partition into cycles of the numbers $(1, 2, ..., n)$ under a random permutation, each possible permutation being given probability $(n!)^{-1}$. As noted below, the distribution arises for a much wider variety of combinatorial objects besides permutations.

2 Distribution, Moments, Structural Properties.

The ESF is most easily described in terms of sequential sampling of animals from an infinite collection of distinguishable species (Fisher *et al* (1943), McCloskey (1965), Engen (1978)). We use this example throughout, except where specific genetic or other properties are discussed. Suppose that the species have (random) frequencies $\boldsymbol{P} = (P_1, P_2, \ldots)$ satisfying

$$0 < P_i < 1, \ i = 1, 2, \ldots, \quad \sum_{i=1}^{\infty} P_i = 1. \tag{1}$$

Let $\eta_1, \eta_2, \ldots$ denote the species of the first, second, ... animal sampled. Conditional on $\boldsymbol{P}$, the η_i are independent and identically distributed, with $\mathrm{P}(\eta_1 = k|\boldsymbol{P}) = P_k, k = 1, 2, \ldots$. The sequence $I_1, I_2, \ldots$ of distinct values observed in $\eta_1, \eta_2, \ldots$ induces a random permutation $\boldsymbol{P}^{\#} = (P_{I_1}, P_{I_2}, \ldots)$ of $\boldsymbol{P}$. The vector $\boldsymbol{P}^{\#}$ is known as the size-biased permutation of $\boldsymbol{P}$.

Consider the sample of n individuals determined by $\eta_1, \ldots, \eta_n$, and write $A_1(n)$ for the number of animals of first species to appear, $A_2(n)$ for the number of animals of the second species to appear, and so on. The number of distinct species to appear in the sample is denoted by K_n. Another way to describe the sample is to record the counts $C_j(n)$, the number of species represented by j animals in the sample. The vector $\boldsymbol{C}_n = (C_1(n), \ldots, C_n(n))$ satisfies $\sum_{j=1}^n jC_j(n) = n$ and $K_n = \sum_{j=1}^n C_j(n)$.

In what follows, we consider the case where $\boldsymbol{P}$ satisfies

$$P_1 = W_1, P_r = (1 - W_1)(1 - W_2) \cdots (1 - W_{r-1})W_r, \ r = 2, 3, \ldots, \tag{2}$$

where

$$W_1, W_2, \ldots \text{ are i.i.d. with density } \theta(1-x)^{\theta-1}, \ \ 0 < x < 1, \ 0 < \theta < \infty. \tag{3}$$

The Ewens Sampling Formula (ESF) gives the distribution of the vector $\boldsymbol{C}_n$ as

$$\mathrm{P}(\boldsymbol{C}_n = \boldsymbol{a}_n) = \frac{n!}{\theta_{(n)}} \prod_{j=1}^n \left(\frac{\theta}{j}\right)^{a_j} \frac{1}{a_j!} \tag{4}$$

where $\theta_{(n)} = \theta(\theta+1)\cdots(\theta+n-1)$ and $\boldsymbol{a}_n = (a_1, a_2, \ldots, a_n)$ is a vector of nonnegative integers satisfying $a_1 + 2a_2 + \cdots + na_n = n$.

The distribution of K_n is (Ewens (1972))

$$\mathrm{P}(K_n = k) = S(n,k)\theta^k/\theta_{(n)}. \tag{5}$$

Here $S(n,k)$ is the coefficient of θ^k in $\theta_{(n)}$, that is, is the absolute value of a Stirling number of the first kind. The distribution of the vector $\boldsymbol{A}_n = (A_1(n), A_2(n), \ldots)$ is determined by (Donnelly and Tavaré (1986))

$$\begin{aligned} &\mathrm{P}(K_n = k, A_i(n) = n_i, i = 1, 2, \ldots, k) \\ &= \frac{\theta^k (n-1)!}{\theta_{(n)} n_k (n_k + n_{k-1}) \cdots (n_k + n_{k-1} + \cdots + n_2)} \cdot \end{aligned} \tag{6}$$

The conditional distribution of $\boldsymbol{C}_n$, given $K_n = k$, is

$$\mathrm{P}(\boldsymbol{C}_n = \boldsymbol{a}_n | K_n = k) = \frac{n!}{S(n,k) \prod_j j^{a_j} a_j!} \cdot \tag{7}$$

An alternative expression for this probability is as follows (Ewens (1972)). Label the K_n species observed in an arbitrary way (independently of the sampling mechanism), and denote the number of animals of species i by N_i, $(i = 1, 2, \ldots, K_n)$. Then

$$\mathrm{P}(N_i = n_i,\ i = 1, \ldots, K_n | K_n = k) = \frac{n!}{k! S(n,k) n_1 \cdots n_k} \tag{8}$$

This conditional distribution is used in the statistical testing of the non-Darwinian theory (see Section 6.1 below).

2.1 Moments and distributions

The joint factorial moments of $\boldsymbol{C}_n$, of arbitrary order, are

$$\mathrm{E} \prod_{j=1}^{n} (C_j(n))_{[r_j]} = \frac{n!}{m!} \frac{\theta_{(m)}}{\theta_{(n)}} \prod_{j=1}^{n} \left(\frac{\theta}{j}\right)^{r_j}, \tag{9}$$

when $m = n - \sum j r_j \geq 0$, and 0 when $m < 0$ (Watterson (1974)); here $x_{[r]} = x(x-1)\cdots(x-r+1)$ for $r = 0, 1, 2, \ldots$.

The number of singleton species is of particular interest. The distribution of this number is

$$\mathrm{P}(C_1(n) = a) = \frac{\theta^a}{a!} \left[\sum_{j=0}^{n-a} (-1)^j \frac{\theta^j}{j!} \frac{(n+1-a-j)_{(a+j)}}{(n+\theta-a-j)_{(a+j)}} \right], \tag{10}$$

so that the mean and the variance of the number of singleton species are, respectively,

$$\frac{n\theta}{n+\theta-1}, \qquad \frac{n(n-1)(n-2+2\theta)\theta}{(n+\theta-2)(n+\theta-1)^2}. \tag{11}$$

It follows from the structure of the urn model in the next Section that

$$K_n = \xi_1 + \xi_2 + \cdots + \xi_n \tag{12}$$

where $\xi_1, \ldots, \xi_n$ are independent Bernoulli random variables with

$$\mathrm{P}(\xi_i = 1) = 1 - \mathrm{P}(\xi_i = 0) = \frac{\theta}{\theta + i - 1}. \tag{13}$$

From this,

$$\mathrm{E}(K_n) = \sum_{i=0}^{n-1} \frac{\theta}{\theta+i}, \quad \mathrm{Var}(K_n) = \sum_{i=1}^{n-1} \frac{\theta i}{(\theta+i)^2} \tag{14}$$

2.2 Urn Models

Now we consider the properties of (4) and (6) for two consecutive sample sizes, n and $n+1$. We denote the history of the sample of size n by $\mathcal{H}_n = (\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots, \boldsymbol{A}_n)$ and ask: Given $\mathcal{H}_n$, what is the conditional probability that the next animal will be of a new species? This probability is found from (4) as

$$\mathrm{P}((n+1)^{th} \text{ animal of a new species} | \mathcal{H}_n) = \frac{\theta}{n+\theta} \cdot \tag{15}$$

The representation (12) follows immediately from this. If a given species has been observed m times ($m > 0$) in the sample of n, the conditional probability that the $(n+1)^{th}$ animal will be of this species is

$$\begin{aligned} &\mathrm{P}((n+1)^{th} \text{ animal of a particular species seen} \\ &\qquad m \text{ times in the sample } |\mathcal{H}_n) = \frac{m}{n+\theta} \cdot \end{aligned} \tag{16}$$

The probabilities (15) and (16) may be used to generate the process $\boldsymbol{A}_n, n = 1, 2 \ldots$ by a sequential urn scheme, starting from $\boldsymbol{A}_1 = (1)$. This model is a special case of an urn scheme of Blackwell and MacQueen (1973) that arises in the context of sampling from a Dirichlet process (see Section 6.2). Hoppe (1984), (1987) exploited a similar urn model in genetics.

2.3 Species Deletion (Non-interference)

Let μ_n denote the distribution of the partition vector $\boldsymbol{C}_n$ when sampling from the species model in (1). We say the sample has the species deletion property if, when an animal is taken at random from the sample, and it is observed that in all there are r animals of this species in the sample, then the partition distribution of the remaining $n-r$ animals is μ_{n-r}. Kingman (1978a,b) shows that the species deletion property holds for the ESF (when μ_n is given by (4)).

2.4 Characterizations

The urn probabilities (15) and (16) and the species deletion property may be used to *characterize* the ESF in the context of sampling from the model (1).

1. (Kingman (1978a,b)). If the species deletion property in Section 2.3 holds, then the vector $\boldsymbol{C}_n$ has distribution μ_n given by the ESF.

2. (The law of succession). Suppose that the sample history $\mathcal{H}_n$ is given. If the conditional probability that the next animal be of a new species depends only on n, then this probability must be of the form $\theta/(\theta+n)$, for some non-negative constant θ (Donnelly (1986)). If, further, the conditional probability that this animal be of a specific species seen m times in the sample depends only on m (the sufficientness principle of Johnson (1932)), then the species partition probability is given by the ESF (Zabell (1995)).

There is a theory of exchangeable random partitions that describes sampling from models slightly more general than (1); see Kingman (1978a), Aldous (1985), and Zabell (1992).

3 Estimation

Equation (4) shows that the ESF is a member of the exponential family of distributions. The complete sufficient statistic for θ is K_n. The maximum likelihood estimator $\hat{\theta}$ is, from (5), given implicitly as the solution of the equation $\sum_{i=0}^{n-1} \hat{\theta}/(\hat{\theta}+i) = K_n$, This estimator is biased, but the bias decreases as n increases. For large n, the variance of $\hat{\theta}$ is $\theta/(\sum_{i=1}^{n-1} i/(\theta+i)^2)$; Ewens (1972).

The only functions of θ admitting unbiased estimation are linear combinations of expressions of the form

$$[(i+\theta)(j+\theta)\cdots(m+\theta)]^{-1} , \tag{17}$$

where $i, j, \ldots, m$ are integers with $1 \le i < j < \cdots < m \le n-1$.

The "law of succession" probability (15) thus does not admit unbiased estimation. However, bounds to unbiased estimation are readily provided by using the inequalities

$$\frac{(n-1)p_n}{n} < \frac{\theta}{n+\theta} < p_n \tag{18}$$

and the MVU estimate $S(n-1, k-1)/S(n,k)$ of p_n.

In genetics one frequently wishes to estimate the homozygosity probability, which in the species context is the probability $(1+\theta)^{-1}$, that two animals taken at random are of the same species. Given $\boldsymbol{C}_n = \boldsymbol{a}_n$, it is natural to estimate this probability by $\sum a_i i(i-1)/n(n-1)$, an estimator occurring often in the genetics literature. The sufficiency of K_n for θ shows, however, that this estimator uses precisely the uninformative part of the data, and that given $K_n = k$, the MVU estimator is $T(n,k)/S(n,k)$, where $T(n,k)$ is the coefficient of θ^k in $\theta(\theta+2)(\theta+3)\cdots(\theta+n-1)$.

4 Relations with Other Distributions

The ESF can be derived from other classical distributions (Watterson (1974)). The first of these is the logarithmic (Johnson *et al* (1992)). Suppose k is fixed and we observe k i.i.d. random variables $N_1, \ldots, N_k$ having the logarithmic distribution $\mathrm{P}(N_1 = j) \propto x^j/j,\ j = 1, 2, \ldots,$ for $0 < x < 1$. Given that $\sum N_i = n$, the distribution of $(N_1, \ldots, N_k)$ is (8). For a second representation, suppose that $Z_1, Z_2, \ldots$ are independent Poisson random variables with $\mathrm{E}(Z_j) = \theta/j$. Then

$$(C_1, \ldots, C_n) =_d (Z_1, \ldots, Z_n | \sum_{j=1}^{n} jZ_j = n), \tag{19}$$

where $=_d$ denotes equality in distribution.

Another representation, called the Feller Coupling, is useful for deriving asymptotic results for the ESF (Arratia *et al* (1992)). Let $\xi_i, i \geq 1$ be independent Bernoulli random variables with distribution (13), and let $C_j(n)$ be the number of spacings of length j between the 1s in the sequence $\xi_1 \xi_2 \ldots \xi_n 1$. Then the distribution of the vector $\boldsymbol{C}_n$ is the ESF. Further, if Z_j is the number of spacings of length j in the infinite sequence $\xi_1 \xi_2 \cdots$, then the Z_j are independent Poisson random variables with mean $\mathrm{E}(Z_j) = \theta/j$.

4.1 The GEM distribution

The distribution of the vector $\boldsymbol{P} = (P_1, P_2, \ldots)$ determined by (2) and (3) is known as the GEM distribution. It is named after McCloskey (1965) and Engen (1978), who introduced it in the context of ecology, and Griffiths (1980), who first noted its genetic importance.

The GEM distribution is a residual allocation model (RAM) (Halmos (1944), Patil and Taillie (1977)), that is, a model of the form (2) where $W_1, W_2, \ldots$ are independent. It is the only RAM $\boldsymbol{P}$ with identically distributed residual fractions for which the size-biased permutation $\boldsymbol{P}^{\#}$ has the same distribution as $\boldsymbol{P}$ (McCloskey (1965), Engen (1975)). For the analog of the non-interference property in Section 2.3 for the GEM see McCloskey (1965) and Hoppe (1986). For further discussion of size-biasing, see Donnelly and Joyce (1989) and Perman *et al* (1992).

The decreasing order statistics $(P_{(1)}, P_{(2)}, \ldots)$ of $\boldsymbol{P}$ have the Poisson-Dirichlet distribution with parameter θ (Kingman (1975)). The GEM is the size-biased permutation of the Poisson-Dirichlet (Patil and Taillie (1977)). For further details about the Poisson-Dirichlet distribution, see Watterson (1976), Ignatov (1982), Tavaré (1987), Griffiths (1988), Kingman (1993), Perman

(1993).

4.2 The Pitman Sampling Formula

The ESF is a particular case of the *Pitman Sampling Formula*, (Pitman (1992), (1995b)), which gives the probability of a species partition $\boldsymbol{C}_n = \boldsymbol{a}_n$ of n animals as

$$\mathrm{P}(\boldsymbol{C}_n = \boldsymbol{a}_n, K_n = k) = \qquad (20)$$
$$\frac{n!}{(\theta+1)_{(n-1)}}\left[(\theta+\alpha)(\theta+2\alpha)\cdots(\theta+(k-1)\alpha)\right]\prod_{j=1}^{n}\left(\frac{(1-\alpha)_{(j-1)}}{j!}\right)^{a_j}\frac{1}{a_j!}.$$

Since we are considering only the infinitely-many-species case, we have the restrictions $0 \le \alpha < 1$, $\theta > -\alpha$. (The other parameter range for which (20) defines a proper distribution is $\alpha = -\kappa, \theta = m\kappa$ for some positive integer m. This corresponds to sampling from a population with m species.) The ESF is then the particular case of the Pitman Sampling Formula when $\alpha = 0$.

The Pitman distribution has several important properties, of which we note here one. Suppose in the RAM model (2), we no longer assume that $W_1, W_2, \ldots$ are identically distributed. Then the most general distribution of W_i for which the distribution of $(P_1, P_2, P_3, \ldots)$ is invariant under size-biased sampling is (Pitman (1995a)) that for which W_i has probability density proportional to $w^{-\alpha}(1-w)^{\theta+i\alpha-1}$. This model for (2) yields the sampling distribution (20). The analog of the Poisson-Dirichlet distribution in the two-parameter setting appears in Pitman and Yor (1995).

5 Approximations

It follows from (10) and the method of moments that random variables $\boldsymbol{C}_n$ with the ESF (4) satisfy, for each fixed b,

$$(C_1(n), \ldots, C_b(n)) \Rightarrow (Z_1, \ldots, Z_b), \qquad (21)$$

as $n \to \infty$, $\Rightarrow$ denoting convergence in distribution. For $\theta = 1$, see Goncharov (1944) and for arbitrary θ see Arratia *et al* (1992). The Feller Coupling may be used to show that the total variation distance between $(C_1(n), \ldots, C_b(n))$ and $(Z_1, \ldots, Z_b)$ is at most $c(\theta)b/n$, where $c(\theta)$ is an explicit constant depending on θ alone. For $\theta \neq 1$, the rate is sharp.

The approximation in (21) covers the case of species represented a small number of times. A functional central limit theorem is available for the number of species represented at most n^t times, for $0 < t \leq 1$ (Hansen (1990)). In particular, the number K_n of species in the sample has asymptotically a Normal distribution with mean and variance $\theta \log n$.

It follows directly from the strong law of large numbers that the proportions $\boldsymbol{A}_n/n$ converge almost surely as $n \to \infty$ to $\boldsymbol{P}^{\#}$, which has the GEM distribution with parameter θ. The decreasing order statistics of $\boldsymbol{A}_n/n$ converge almost surely to the Poisson-Dirichlet distribution with parameter θ (Kingman (1975)).

6 Applications

6.1 Genetics

The original aim in devising (4) was to obtain a testing procedure for the non-Darwinian theory, since (4) provides the null hypothesis distribution for this theory. The parameter θ depends, in this context, on an unknown mutation parameter, an unknown population size and on unknown details about the evolutionary model. However the conditional distribution (9) does not depend on θ and hence may be used as an objective basis for a test of the non-Darwinian theory. Watterson (1978) shows that a suitable test statistic is $\sum a_i i^2/n^2$ and provides various examples of the application of this at different gene loci. Anderson (see Ewens (1979), Appendix C) provides charts allowing rapid testing.

The ESF was derived directly by a genealogical argument by Karlin and McGregor (1972). The Poisson-Dirichlet distribution arises as the stationary distribution of the ranked allele frequencies in the infinitely-many-alleles model (Watterson (1976)). Equation (6) provides the distribution of alleles frequencies when the alleles are ordered by decreasing age (Donnelly and Tavaré (1986)), and this provides significant evolutionary information. See also Kelly (1979, Chapter 7). Correspondingly, the GEM distribution is the stationary distribution of the infinitely-many-alleles model when the types are ordered by age (Griffiths (1980)). The ESF may also be derived directly as a consequence of mutation in the coalescent (Kingman (1982a,b,c)). See also Hoppe (1987), Ewens (1990).

6.2 Bayesian statistics

Dirichlet processes on a set S (Ferguson (1973)) are often used as priors over spaces of probability distributions on S. Suppose that the measure α of the process is non-atomic, set $\theta = \alpha(S) < \infty$, let $\boldsymbol{P} = (P_1, P_2, \ldots)$ have the GEM distribution with parameter θ and let $X_1, X_2, \ldots$ be i.i.d. random elements of S with distribution $\alpha(\cdot)/\theta$, independent of $\boldsymbol{P}$. Sethuraman and Tiwari (1982) represent the Dirichlet process as atoms of height P_i at locations $X_i,\ i = 1, 2, \ldots$. A similar representation arises as the stationary distribution of the infinitely-many-alleles measure-valued diffusion in population genetics (Ethier and Kurtz (1994)). Thus the Bayesian setting is essentially the same as sampling animals from a GEM population where the labels (determined by the X_i) of the animals are recorded as well. Antoniak (1974) showed that the ESF gives the distribution of the partition induced by a sample from a Dirichlet process. See Ferguson *et al* (1992) and Sethuraman (1994) for recent developments.

6.3 Permutations

A permutation of the integers $1, 2, \ldots, n$ may be decomposed into an ordered product of cycles by beginning the first cycle with the integer 1, the second with the smallest integer not in the first cycle, and so on. For any $\theta > 0$, a random permutation, decomposed in this way, may be generated by Dubins and Pitman's *Chinese restaurant process* (cf. Aldous (1985)): Integer 1 begins the first cycle. With probability $\theta/(\theta+1)$ integer 2 starts the second cycle, and with probability $1/(\theta+1)$ it joins the first cycle, to the right of 1. Once the first $r-1$ integers have been placed in cycles, integer r starts a new cycle with probability $\theta/(\theta+r-1)$, or is placed in an existing cycle, to the right of a random chosen one of $1,2,\ldots,r-1$. After n steps of this process, the probability of obtaining a particular permutation π with k cycles is $\theta^k/\theta_{(n)}$. Since the number of n-permutations having a_i cycles of size i is $n!/\prod j^{a_j} a_j!$, it follows that the joint distribution of the numbers C_j of cycles of size j, $j = 1, 2, \ldots, n$ is given by the ESF (4).

The case $\theta = 1$ corresponds to random permutations, which have been widely studied in the literature. Shepp and Lloyd (1966) show that the proportions in the largest, second largest, $\ldots$ asymptotically as $n \to \infty$ the Poisson-Dirichlet distribution. Erdős and Turán (1967) showed that the logarithm of the order (the least common multiple of its cycle lengths) of such a random permutation has asymptotically a Normal distribution with mean $\log^2 n/2$ and variance $\log^3 n/3$. See also Vershik and Shmidt (1977) for a connection with the GEM distribution. The functional central limit theorem for the medium-sized cycles is given by DeLaurentis and Pittel (1985). When $\theta = 1$, random permutations are intimately connected to the theory of records

(Ignatov (1981), Goldie (1989)).

For arbitrary θ, equation (6) describes the joint distribution of the ordered cycle lengths. It follows that asymptotically the proportions in these cycles have the GEM distribution. Other approximations follow directly from Section 5. For the Erdős-Turán law for arbitrary θ, see Barbour and Tavaré (1994).

6.4 Ecology

In ecology, a long-standing problem concerned the species allocation of animals when species do not interact, in the sense that removal of one species does not affect the relative abundances of other species. Several attempts in the ecological literature, notably the "broken stick" model of MacArthur (1957), attempted to resolve this question. The non-interference property of the ESF shows that this distribution provides the required partition, and (4) has been applied in various ecological contexts (Caswell (1976), Lambshead (1986), Lambshead and Platt (1985)) where non-interference can be assumed.

The description of species diversity, through conditioned or unconditioned logarithmic distributions, has a long history in ecology (Fisher (1943), Fisher *et al* (1943), McCloskey (1965), Engen (1975)). For a summary see Watterson (1974).

6.5 Physics

The urn representation (15) and (16) is related to an urn representation of three classical partition formulae in physics, (where a ball represents a particle and an urn a "cell", or energy level). Constantini (1987) considers the case where balls are placed sequentially into a collection of m urns so that, if among the first n balls there are n_j in urn j, the probability that ball $n+1$ is placed in this urn is

$$\frac{n_j + \delta}{n + m\delta} \tag{22}$$

for some constant δ. The Maxwell-Boltzmann, Bose-Einstein and Fermi-Dirac statistics follow when $\delta \to \infty$, $\delta = 1$, $\delta = -1$ respectively, while (15) and (16) show that the ESF follows when $\delta \to 0$, $m \to \infty$ with $m\delta = \theta$. See also Keener *et al* (1987).

None of the physics partition formulae satisfy the non-interference property. Direct application of the ESF in physics, in cases where the non-interference property is required, are given by Sibuya *et al* (1990), Mekjian (1991), Mekjian and Lee (1991), and Higgs (1995).

6.6 The spread of news and rumors

Bartholomew (1973) describes a simple model of the spread of news (or a rumor) throughout a population of n individuals. It is supposed that there is a source (e.g. a radio station) broadcasting the news, and that each person in the population first hears the news either from the source or from some other individual. A person not knowing the news hears it from the source at rate α, and from a person who has heard the news at rate β. The analogy with (15) and (16) is apparent, and Bartholomew shows that, when all persons in the population have heard the news, the probability that k heard it directly from the source is given by (5), with $\theta = \alpha/\beta$.

This model is a Yule process with immigration and much more can be said. Individuals can be grouped into components, each consisting of exactly one person who first heard the news from the source, together with those individuals who first heard the news through some chain of individuals deriving from this person. Joyce and Tavaré's (1987) analysis applies directly to show among other things that the joint distribution of the component sizes is given by the ESF.

6.7 The law of succession

The law of succession problem is perhaps the most classical in all of probability theory. In the sampling of species context, we ask, given a sample of n animals, for the probability that animal $n+1$ is of a previously unobserved species and also for the probability that this animal is of a species seen m (> 0) times in the sample.

Clearly further assumptions are necessary to obtain concrete answers. For simplicity, we continue in the setting of (1) and we assume the sufficientness postulate. If we assume also that the probability that animal $n+1$ is of a new species depends only on n and the number k of species seen in the sample, then (Pitman (1995b)), Zabell (1995)) the species partition in the sample must be given by Pitman Sampling Formula (20). This implies that the probability that animal $n+1$ is of a previously unobserved species is $(k\alpha+\theta)/(n+\theta)$, and that it is of a particular species seen m times in the sample is $(m-\alpha)/(n+\theta)$, where $0 \leq \alpha < 1$, $\theta > -\alpha$. This remarkable result represents the most significant recent advance in the theory of the law of succession. If we further require the probability that animal $n+1$ be of a new species depends only on n, then

$\alpha = 0$ and the species probability structure of the sample reduces to the ESF.

6.8 Prime numbers

Let N be an integer drawn at random from the set $1, 2, \ldots, n$, and write $N = p_1 p_2 p_3 \cdots$, where $p_1 \geq p_2 \geq p_3 \cdots$ are the prime factors of N. Writing $L_i = \log p_i / \log N,\ i \geq 1$, Billingsley (1972) showed that $(L_1, L_2, \ldots)$ has asymptotically as $n \to \infty$ the Poisson-Dirichlet distribution with parameter $\theta = 1$. Dickman (1930) studied the distribution of L_1. Donnelly and Grimmett (1993) provide an elementary proof using size-biasing and the GEM distribution.

6.9 Random mappings

The partition probability (4) appears also in the field of random mappings. Suppose random mapping of $(1, 2, \ldots, N)$ to $(1, 2, \ldots, N)$ is made, each mapping having probability N^{-N}. Any mapping defines a number of *components*, where i and j are in the same component if some functional iterate of i is identical to some functional iterate of j. In the limit $N \to \infty$, the normalized component sizes have a Poisson-Dirichlet distribution with $\theta = 1/2$ (Aldous (1985)), and the images of the components in the set $\{1, 2, \ldots, n\}$, for any fixed n, have the distribution (4), again with $\theta = 1/2$ (Kingman (1977)).

6.10 Combinatorial structures

The joint distribution of the component counting process of many decomposable combinatorial structures satisfies the relation (19) for appropriate independent random variables Z_i (Arratia and Tavaré (1994)). Examples include random mappings (discussed in the last section), factorization of polynomials over a finite field, and forests of labeled trees. When $i\mathrm{E}(Z_i) \to \theta,\ i\mathrm{P}(Z_i = 1) \to \theta$ for some $\theta \in (0, \infty)$ as $i \to \infty$, the counts of large components are close, in total variation distance, to the corresponding counts for the ESF with parameter θ (Arratia *et al* (1995)). Polynomial factorization satisfies $\theta = 1$. Poisson-Dirichlet approximations for a related class of combinatorial models are given by Hansen (1994).

Bibliography

Aldous, D. J. (1985) Exchangeability and related topics, in: École d'été de probabilités de Saint-Flour XIII-1983 (P. L. Hennequin, éd.), *Lecture Notes in Mathematics 1117, Springer-Verlag, Berlin*, 2-198.

Antoniak, C. E. (1974) Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics*, **2**, 1152-1174.

Arratia, R. A. and Tavaré, S. (1994) Independent process approximations for random combinatorial structures. *Advances in Mathematics*, **104**, 90-154.

Arratia, R. A., Barbour, A. D. and Tavaré, S. (1992) Poisson process approximations for the Ewens sampling formula. *Annals of Applied Probability*, **2**, 519-535.

Arratia, R. A., Barbour, A. D. and Tavaré, S. (1995) Logarithmic combinatorial structures. In preparation.

Barbour, A. D. and Tavaré, S. (1994) A rate for the Erdős-Turán law. *Combinatorics, Probability and Computing*, **3**, 167-176.

Bartholomew, D. J. (1973) *Stochastic Models for Social Processes*, Second Edition, London: Wiley.

Billingsley, P. (1972) On the distribution of large prime divisors. *Periodica Mathematica Hungarica*, **2**, 283-289.

Blackwell, D. and MacQueen, J. B. (1973) Ferguson distributions via Pólya urn schemes. *Annals of Statistics*, **1**, 353-355.

Caswell, H. (1976) Community structure: a neutral model analysis. *Ecological Monographs*, **46**, 327-353.

Constantini, D. (1987) Symmetry and distinguishability of classical particles. *Physics Letters A*, **123** 433-436.

DeLaurentis, J.M. and Pittel, B. (1985) Random permutations and Brownian motion. *Pacific J. Math.*, **119**, 287-301.

Dickman, K. (1930) On the frequency of numbers containing prime factors of a certain relative magnitude. *Arkiv for Matematik, Astronomi och Fysik*, **22**, 1-14.

Donnelly, P. (1986) Partition structures, Polya urns, the Ewens sampling formula, and the ages of alleles. *Theoretical Population Biology*, **30**, 271-288.

Donnelly, P. and Grimmett, G. (1993) On the asymptotic distribution of large prime factors. *Journal of the London Mathematical Society*, **47**, 395-404.

Donnelly, P. and Joyce, P. (1989) Continuity and weak convergence of ranked and size-biased permutations on an infinite simplex. *Stochastic Processes and their Applications*, **31**, 89-103.

Donnelly, P. and Tavaré, S. (1986) The ages of alleles and a coalescent. *Advances in Applied Probability*, **18**, 1-19.

Engen, S. (1975) A note on the geometric series as a species frequency model. *Biometrika*, **62**, 697-699.

Engen, S. (1978) *Stochastic Abundance Models with Emphasis on Biological Communities and Species Diversity.* London: Chapman and Hall.

Erdős, P. and Turán, P. (1967) On some problems of a statistical group theory. III. *Acta Mathematica Academiae Scientiarum Hungaricae*, **18**, 309-320.

Ethier, S. N. and Kurtz, T. G. (1994) Convergence to Fleming-Viot processes in the weak atomic topology. *Stochastic Processes and their Applications*, **54**, 1-27.

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

Ewens, W. J. (1979) *Mathematical Population Genetics*, Berlin: Springer-Verlag.

Ewens W. J. (1990) Population genetics theory – the past and the future, in: *Mathematical and Statistical Developments of Evolutionary Theory,* (S. Lessard, ed.), pp. 177-227. Kluwer Dordrecht, Amsterdam.

Ferguson, T. S. (1973) A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, **1**, 209-230.

Ferguson, T. S. , Phadia, E. G. and Tiwari, R. C. (1992) Bayesian nonparametric inference, in: *Current Issues in Statistical Inference: Essays in Honor of D. Basu*, (M. Ghosh and P. K. Patnak, eds). IMS Lecture Notes-Monograph Series, **17**, 127-150.

Fisher, R. A. (1943) A theoretical distribution for the apparent abundance of different species. *Journal of Animal Ecology*, **12**, 54-57.

Fisher, R. A., Corbet, A. S., and Williams, C. B. (1943) The relation between the number of species and the number of individuals in a random sample from an animal population. *Journal of Animal Ecology*, **12**, 42-58.

Goldie, C. M. (1989) Records, permutations and greatest convex minorants. *Mathematical Proceedings of the Cambridge Philosophical Society*, **106**, 169-177.

Goncharov, V.L. (1944) Some facts from combinatorics. *Izvestia Akad. Nauk. SSSR, Ser. Mat.*, **8**, 3-48. See also: On the field of combinatory analysis. (1962) *Translations of the American Mathematical Society*, **19**, 1-46.

Griffiths, R. C. (1980) Unpublished notes.

Griffiths, R. C. (1988) On the distribution of points in a Poisson-Dirichlet process. *Journal of Applied Probability*, **25**, 336-345.

Halmos, P. R. (1944) Random alms. *Annals of Mathematical Statistics*, **15**, 182-189.

Hansen, J.C. (1990) A functional central limit theorem for the Ewens Sampling Formula. *Journal of Applied Probability*, **27**, 28-43.

Hansen, J. C. (1994) Order statistics for decomposable combinatorial structures. *Random Structures and Algorithms*, **5**, 517-533.

Higgs, P. G. (1995) Frequency distributions in population genetics parallel those in statistical physics. *Physical Review E*, **51**, 95-101.

Hoppe, F. M. (1984) Polya-like urns and the Ewens sampling formula. *Journal of Mathematical Biology*, **20**, 91-99.

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

Hoppe, F. M. (1987) The sampling theory of neutral alleles and an urn model in population genetics. *Journal of Mathematical Biology*, **25**, 123-159.

Ignatov, T. (1981) Point processes generated by order statistics and their applications, in: *Point Processes and Queuing Problems*, (P. Bartfái and J. Tomkó, eds.), 109-116: North Holland.

Ignatov, T. (1982) On a constant arising in the asymptotic theory of symmetric groups, and on Poisson-Dirichlet measures. (1982). *Theory of Probability and its Applications*, **27**, 136-147.

Johnson, W. E. (1932) *Logic, Part III: The Logical Foundations of Science*, Cambridge: Cambridge University Press.

Johnson, N. L., Kotz, S. and Kemp, A. W. (1992) *Univariate Discrete Distributions*, New York: Wiley.

Joyce, P. and Tavaré, S. (1987) Cycles, permutations and the structure of the Yule process with immigration. *Stochastic Processes and their Applications*, **25**, 309-314.

Karlin, S. and McGregor, J. (1972) Addendum to a paper of W. Ewens. *Theoretical Population Biology*, **3**, 113-116.

Keener, R., Rothman, E. and Starr, N. (1987) Distributions of partitions. *Annals of Statistics*, **15**, 1466-1481.

Kelly, F.P. (1979) *Reversibility and Stochastic Networks*: Wiley, New York.

Kingman, J. F. C. (1975) Random discrete distributions. *Journal of the Royal Statistical Society B*, **37**, 1-22.

Kingman, J. F. C. (1977) The population structure associated with the Ewens sampling formula. *Theoretical Population Biology*, **11**, 274-283.

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

Kingman, J. F. C. (1978b) The representation of partition structures. *Journal of the London Mathematical Society*, **18**, 374-380.

Kingman, J. F. C. (1980) *Mathematics of Genetic Diversity*, Philadelphia: SIAM.

Kingman, J. F. C. (1982a) On the genealogy of large populations. *Journal of Applied Probability*, **19A**, 27-43.

Kingman, J. F. C. (1982b) The coalescent. *Stochastic Processes and their Applications,* **13**, 235-248.

Kingman, J. F. C. (1982c) Exchangeability and the evolution of large populations, in *Exchangeability in Probability and Statistics*, (G. Koch, F. Spizzichino, eds), pp. 97-112. North-Holland Publishing Company

Kingman, J. F. C. (1993) *Poisson Processes.* Oxford: Clarendon Press.

Lambshead, P. J. D. (1986) Sub-catastrophic sewage and industrial waste contamination as revealed by marine nematode faunal analysis. *Marine Ecology Progress Series*, **29**, 247-260.

Lambshead, P. J. D. and Platt, H. M. (1985) Structural patterns of marine benthic assemblages and their relationship with empirical statistical models, in: Nineteenth European Marine Biology Symposium. *Cambridge University Press*, Cambridge, 371-380.

MacArthur, R. H. (1957) On the relative abundance of bird species. *Proceedings of the National Academy of Sciences, USA*, **43**, 293-295.

McCloskey, J. W. (1965) A model for the distribution of individuals by species in an environment. Unpublished PhD Thesis, Michigan State University.

Mekjian, A. Z. (1991) Cluster distributions in physics and genetic diversity. *Physical Review A*, **44**, 8361-8374.

Mekjian, A. Z. and Lee, S. J. (1991) Models of fragmentation and partitioning phenomena based on the symmetric group S_n and combinatorial analysis. *Physical Review A*, **44**, 6294-6311.

Patil, G. P. and Taillie, C. (1977) Diversity as a concept and its implications for random communities. *Bulletin of the International Statistical Institute*, *XLVII*, 497-515.

Perman, M. (1993) Order statistics for jumps of normalized subordinators. *Stochastic Processes and their Applications*, **46**, 267-281.

Perman, M., Pitman, J., and Yor, M. (1992) Size-biased sampling of Poisson point processes and excursions. *Probability Theory and Related Fields*, **92**, 21-39.

Pitman, J. (1992) The two-parameter generalization of Ewens' random partition structure. Technical Report #345, Department of Statistics, University of California at Berkeley.

Pitman, J. (1995a) Random discrete distributions invariant under size-biased permutation. *Journal of Applied Probability.* (To appear).

Pitman, J. (1995b) Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields*, **102**, 145-158.

Pitman, J. and Yor, M. (1995) The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. Technical Report #427, Department of Statistics, University of California at Berkeley.

Sethuraman, J. (1994) A constructive definition of Dirichlet priors. *Academica Sinica*, **4**, 639-650.

Sethuraman, J. and Tiwari, R. C. (1982) Convergence of Dirichlet measures and the interpretation of their parameter, in *Statistical Decision Theory and Related Topics III 2*, (Gupta and Berger, eds), 305-315, New York: Academic Press.

Shepp, L. A. and Lloyd, S. P. (1966) Ordered cycle lengths in a random permutation. *Transactions of the American Mathematical Society*, **121**, 340-357.

Sibuya, M., Kawai, T., and Shida, K. (1990) Equipartition of particles forming clusters by inelastic collisions. *Physica A*, 167, 676-689.

Tavaré, S. (1987) The birth process with immigration and the genealogical structure of large populations. *Journal of Mathematical Biology*, **25**, 161-168.

Vershik, A. M. and Shmidt, A. A. (1977) Limit measures arising in the asymptotic theory of symmetric groups I. *Theory of Probability and its Applications*, **22**, 70-85.

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

Watterson, G. A. (1976) The stationary distribution of the infinitely-many neutral alleles diffusion model. *Journal of Applied Probability*, **13**, 639-651.

Watterson, G. A. (1978) The homozygosity test of neutrality. *Genetics*, **88**, 405-417.

Zabell, S. L. (1992) Predicting the unpredictable. *Synthese*, **90**, 205-232.

Zabell, S. L. (1995) The continuum of inductive methods revisited. *Pittsburgh-Konstanz series in the History and Philosophy of Science.* (To appear).