CHAPTER 41

Multivariate Ewens Distribution¹

1 GENESIS AND HISTORY

The Multivariate Ewens Distribution (MED), 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 discovered independently 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 MED 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 \( \theta \), which in the genetic context is related to (a) the rate of mutation of the genes to new allelic types, (b) the population size, and (c) 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 \( 1, 2, \ldots, n \) 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.

¹We thank Professors S. Tavaré and W. J. Ewens for providing us an original write-up of this chapter, and we thank J. W. Pitman for comments on early drafts. The distribution described in this chapter, which originated from applications in genetics and also independently in Bayesian statistical methodology, serves as a striking example of adaptability and universality of statistical methodology for scientific explorations in various seemingly unrelated fields.
2 DISTRIBUTION, MOMENTS, AND STRUCTURAL PROPERTIES

The MED is most easily described in terms of sequential sampling of animals from an infinite collection of distinguishable species [Fisher, Corbet, and Williams (1943), McCloskey (1965), and Engen (1978)]. We use this example throughout, except where specific genetic or other properties are discussed. Suppose that the species have (random) frequencies \( P = (P_1, P_2, \ldots) \) satisfying

\[
0 < P_i < 1, \quad i = 1, 2, \ldots, \sum_{i=1}^{\infty} P_i = 1.
\] (41.1)

Let \( \eta_1, \eta_2, \ldots \) denote the species of the first, second, \ldots, animal sampled. Conditional on \( P \), the \( \eta_i \) are independent and identically distributed, with \( \Pr[\eta_i = k \mid P] = P_k \), \( k = 1, 2, \ldots \). The sequence \( l_1, l_2, \ldots \) of distinct values observed in \( \eta_1, \eta_2, \ldots \) induces a random permutation \( P^* = (P_1, P_2, \ldots) \) of \( P \). The vector \( P^* \) is known as the size-biased permutation of \( P \).

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 \( C(n) = (C_1(n), \ldots, C_n(n)) \) 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 \( P \) satisfies

\[
P_1 = W_1, \quad P_r = (1 - W_1)(1 - W_2) \cdots (1 - W_{r-1}) W_r, \quad r = 2, 3, \ldots,
\] (41.2)

where, for some \( 0 < \theta < \infty \),

\[
W_1, W_2, \ldots \text{ are i.i.d. with density } \theta(1 - x)^{\theta-1}, \quad 0 < x < 1.
\] (41.3)

The MED gives the distribution of the vector \( C(n) \) as

\[
\Pr[C(n) = a(n)] = \frac{n!}{\theta^{\lvert a \rvert}} \prod_{j} \frac{(\theta/j)^{a_j}}{a_j!},
\] (41.4)

where, as earlier, \( \theta^{\lvert a \rvert} = \theta(\theta + 1) \cdots (\theta + n - 1) \) and \( a(n) = (a_1, a_2, \ldots, a_n) \) is a vector of non-negative integers satisfying \( a_1 + 2a_2 + \cdots + na_n = n \).

The distribution of \( K_n \) is [Ewens (1972)]

\[
\Pr[K_n = k] = \bar{s}(n, k) \theta^k / \theta^{\lvert a \rvert}. \tag{41.5}
\]

Here \( \bar{s}(n, k) \) is the coefficient of \( \theta^k \) in \( \theta^{\lvert a \rvert} \)—that is, a Stirling number of the third kind (see Chapter 34). The distribution of the vector \( A(n) = (A_1(n), A_2(n), \ldots) \) is
determined by [Donnelly and Tavaré (1986)]

\[
Pr[K_n = k, A_i(n) = n_i, i = 1, 2, \ldots, k] = \frac{\theta(n - 1)!}{\theta(n) n_k (n_k + n_{k-1}) \cdots (n_k + n_1 + \cdots + n_2)},
\]

(41.6)

for \( n_1 + \cdots + n_k = n \).

The conditional distribution of \( C(n) \), given \( K_n = k \), is

\[
Pr[C(n) = s(n) \mid K_n = k] = \frac{n!}{s(n, k) \prod_j j^{a_j} \theta^{a_j}}.
\]

(41.7)

An alternative expression for this probability is as follows [due to 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

\[
Pr[N_i = n_i, i = 1, \ldots, K_n \mid K_n = k] = \frac{n!}{k! \theta(n, k) n_1 \cdots n_k}.
\]

(41.8)

This conditional distribution is used in the statistical testing of the non-Darwinian theory (see Section 6.1 on page 239).

2.1 Moments

The joint factorial moments of \( C(n) \), of arbitrary order, are

\[
E \left[ \prod_{j=1}^{n} (C_j(n))^r_j \right] = \frac{n! \theta^{[m]} \prod_j \theta^{a_j}}{m! \theta^{[n]} \prod_j (\theta^{-1})^{r_j}}
\]

(41.9)

when \( m = n - \sum j r_j \geq 0 \) and are 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

\[
Pr[C_1(n) = a] = \frac{\theta^a}{a!} \left[ \sum_{j=0}^{n-a} (-1)^j j^a (n + 1 - a - j)^{[a+j]} \right],
\]

(41.10)

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

\[
\frac{n \theta}{n + \theta - 1}, \quad \frac{n(n - 1)(n - 2 + 2 \theta) \theta}{(n - \theta - 2)(n + \theta - 1)^2}.
\]

(41.11)

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

\[
K_n = \xi_1 + \xi_2 + \cdots + \xi_n.
\]

(41.12)
where \( \xi_1, \ldots, \xi_n \) are independent Bernoulli random variables with

\[
\Pr[\xi_i = 1] = 1 - \Pr[\xi_i = 0] = \frac{\theta}{\theta + i - 1}.
\]

From this [for example, Cauchy (1905)],

\[
E[K_n] = \sum_{i=0}^{n-1} \frac{\theta}{\theta + i}, \quad \text{var}(K_n) = \sum_{i=1}^{n-1} \frac{\theta i}{(\theta + i)^2}.
\]

2.2 Urn Models

Now we consider the properties of (41.4) and (41.6) for two consecutive sample sizes, \( n \) and \( n + 1 \). We denote the history of the sample of size \( n \) by \( \mathcal{H}_n = (A(1), A(2), \ldots, A(n)) \) 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 (41.4) as

\[
\Pr[(n + 1)\text{th animal of a new species } | \mathcal{H}_n] = \frac{\theta}{n + \theta}.
\]

The representation (41.12) 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 \( (n + 1) \text{th animal will be of this species} \) is

\[
\Pr[(n + 1)\text{th animal of a particular species seen } m \text{ times in the sample } | \mathcal{H}_n] = \frac{m}{n + \theta}.
\]

The probabilities (41.15) and (41.16) may be used to generate the process \( A(n), n = 1, 2 \ldots \) by a sequential urn scheme, starting from \( A(1) = 1 \). 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 (Noninterference)

Let \( \mu_n \) denote the distribution of the partition vector \( C(n) \) when sampling from the species model in (41.1). 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 \( \mu_{n-r} \). Kingman (1978a,b) shows that the species deletion property holds for the MED [when \( \mu_n \) is given by (41.4)].

2.4 Characterizations

The urn probabilities (41.15) and (41.16) and the species deletion property may be used to characterize the MED in the context of sampling from the model (41.1).
i. If the species deletion property in Section 2.3 holds, then the vector $C(n)$ has distribution $\mu_n$ given by the ESF [Kingman (1978a,b)].

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 $\theta$ [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 MED [Zabell (1996)].

There is a theory of exchangeable random partitions that describes sampling from models slightly more general than (41.1); see Kingman (1978a), Aldous (1985), and Zabell (1992).

3 ESTIMATION

Equation (41.4) shows that the MED is a member of the exponential family of distributions: see, for example, Chapter 34. The complete sufficient statistic for $\theta$ is $K_n$. The maximum likelihood estimator $\hat{\theta}$ is, from (41.5), 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 $\theta$ admitting unbiased estimation are linear combinations of expressions of the form

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

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

The “law of succession” probability (41.15) thus does not admit unbiased estimation. However, bounds to unbiased estimation are readily provided by using the inequalities

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

and the MVU estimate $\hat{n}(n - 1, k - 1)/\hat{n}(n, k)$ 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 $C(n) = a(n)$, it is natural to estimate this probability by $\sum a_i(i - 1)/n(n - 1)$, an estimator occurring often in the genetics literature. The sufficiency of $K_n$ for $\theta$ shows, however, that this estimator uses precisely the uninformative part of the data and that, given $K_n = k$, the MVU estimator is $T(n, k)/\hat{n}(n, k)$, where $T(n, k)$ is the coefficient of $\theta^k$ in $\theta(\theta + 2)(\theta + 3) \cdots (\theta + n - 1)$. 


4 RELATIONS WITH OTHER DISTRIBUTIONS

The MED can be derived from other classical distributions [Watterson (1974)]. The first of these is the logarithmic (see, for example, Chapter 8). Suppose $k$ is fixed and we observe $k$ i.i.d. random variables $N_1, \ldots, N_k$ having the logarithmic distribution $\Pr[N_i = j] \propto x^j / j$, $j = 1, 2, \ldots$, for $0 < x < 1$. Given that $\sum N_i = n$, the distribution of $(N_1, \ldots, N_k)$ is (41.8). For a second representation, suppose that $Z_1, Z_2, \ldots$ are independent Poisson random variables with $\Pr[Z_j = \theta / j]$. Then

$$(C_1, \ldots, C_n)' \overset{d}{=} \left( Z_1, \ldots, Z_n \left| \sum_{j=1}^n j Z_j = n \right. \right)' , \quad (41.19)$$

where $\overset{d}{=} \text{denotes equality in distribution.}

Another representation, called the Feller Coupling, is useful for deriving asymptotic results for the MED [Arratia, Barbour, and Tavaré (1992)]. Let $\xi_i, i \geq 1$, be independent Bernoulli random variables with distribution (41.13), and let $C_j(n)$ be the number of spacings of length $j$ between the 1s in the sequence $\xi_1 \xi_2 \cdots \xi_n$. Then the distribution of the vector $C(n)$ is the MED. 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 $\Pr[Z_j = \theta / j]$.

4.1 The GEM Distribution

The distribution of the vector $P = (P_1, P_2, \ldots)$ determined by (41.2) and (41.3) is known as the GEM distribution (Generalized Engen–McCloskey 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 (41.2) where $W_1, W_2, \ldots$ are independent. It is the only RAM $P$ with identically distributed residual fractions for which the size-biased permutation $P^*$ has the same distribution as $P$ [McCloskey (1965), Engen (1975)]. For the analog of the noninterference 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), Perman, Pitman, and Yor (1992), and Chapters 3 (p. 146) and 43 (Section 5).


4.2 The Pitman Sampling Formula

The MED is a particular case of the Pitman Sampling Formula [Pitman (1992, 1995)], which gives the probability of a species partition $C(n) = a(n)$ of $n$ animals as
\[ P_1[\mathcal{C}(n) = a(n), K_n = k] = \frac{n!}{(\theta + 1)^{n-1}} \left[ (\theta + \alpha)(\theta + 2\alpha) \cdots (\theta + (k-1)\alpha) \right] \times \prod_{j=1}^{n} \left( \frac{(1-\alpha)^{j-1}}{j!} \right)^{a_j} \frac{1}{a_j!} \] (41.20)

Since we are considering only the infinitely many species case, we have the restrictions \(0 \leq \alpha < 1, \theta > -\alpha\). The other parameter range for which (41.20) 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 MED 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 (41.2) 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 [Pitman (1996)] is that for which \(W_i\) has probability density proportional to \(w^{-a}(1-w)^{\theta+a-1}\). This model for (41.2) yields the sampling distribution (41.20). The analogue of the Poisson-Dirichlet distribution in the two-parameter setting appears in Pitman and Yor (1995).

5 APPROXIMATIONS

It follows from (41.10) and the method of moments that random variables \(C(n)\) with the MED (41.4) satisfy, for each fixed \(b\),

\[(C_1(n), \ldots, C_b(n))' \Rightarrow (Z_1, \ldots, Z_b)',\] (41.21)

as \(n \to \infty\), \(\Rightarrow\) denoting convergence in distribution. For \(\theta = 1\) see Goncharov (1944), and for arbitrary \(\theta\) see Arratia, Barbour, and Tavaré (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 \(\theta\) alone. For \(\theta \neq 1\), the rate is sharp.

The approximation in (41.21) 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 \(n'\) 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 \(A(n)/n\) converge almost surely as \(n \to \infty\) to \(P^x\), which has the GEM distribution with parameter \(\theta\). The decreasing order statistics of \(A(n)/n\) converge almost surely to the Poisson-Dirichlet distribution with parameter \(\theta\) [Kingman (1975)].
6 APPLICATIONS

6.1 Genetics

The original aim in devising (41.4) was to obtain a testing procedure for the non-Darwinian theory, since (41.4) provides the null hypothesis distribution for this theory. The parameter $\theta$ depends, in this context, on an unknown mutation parameter, an unknown population size, and unknown details about the evolutionary model. However, the conditional distribution (41.9) does not depend on $\theta$ 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 MED 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 (41.6) provides the distribution of allele 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 MED may also be derived directly as a consequence of mutation in the coalescent [Kingman (1980, 1982a–c)]. See also Hoppe (1987) and 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 $\alpha$ of the process is nonatomic, and assume $\theta = \alpha(S) < \infty$. Let $P = (P_1, P_2, \ldots)$ have the GEM distribution with parameter $\theta$ and let $X_1, X_2, \ldots$ be i.i.d. random elements of $S$ with distribution $\alpha(\cdot)/\theta$, independent of $P$. Sethuraman and Tiwari (1981) 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 MED gives the distribution of the partition induced by a sample from a Dirichlet process. See Ferguson, Phadia, and Tiwari (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 $\pi$ with $k$ cycles is $\theta^k/ \theta^{[n]}$. Since the number of $n$-permutations having $a_i$ cycles of size $i$ is $n!/ \prod j^n a_j!$, 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 MED (41.4).

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, ... cycle lengths, asymptotically as $n \rightarrow \infty$, have a limiting 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) and Goldie (1989)].

For arbitrary $\theta$, Eq. (41.6) 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 $\theta$, see Barbour and Tavare (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 noninterference property of the MED shows that this distribution provides the required partition, and (41.4) has been applied in various ecological contexts [Caswell (1976), Lambshead (1986), and 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, Corbet, and Williams (1943), McCloskey (1965), Engen (1975), and Chapter 8 of Johnson, Kotz, and Kemp (1992)]. For a summary, see Watterson (1974).

6.5 Physics

The urn representation (41.15) and (41.16) is related to an urn representation of three classical partition formulae in physics [Bose–Einstein, Fermi–Dirac, and Maxwell–Boltzmann; for details see Johnson and Kotz (1977)] where a ball represents a “particle” and an urn represents 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

\[
\frac{n_j + \delta}{n + m\delta}
\]  

(41.22)

for some constant \( \delta \). The Maxwell–Boltzmann, Bose–Einstein and Fermi–Dirac statistics follow when \( \delta \rightarrow \infty, \delta = 1, \delta = -1 \) respectively, while (41.15) and (41.16) show that the MED follows when \( \delta \rightarrow 0, m \rightarrow \infty \) with \( m\delta = \theta \). See also Keener, Rothman, and Starr (1987).

None of the physics partition formulae satisfy the noninterference property. Direct application of the MED in physics, in cases where the noninterference property is required, are given by Sibuya, Kawai, and Shida (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 \( \alpha \), as well as from a person who has heard the news at rate \( \beta \). The analogy with (41.15) and (41.16) 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 (41.5), with \( \theta = \alpha/\beta \).

This model is a Yule process with immigration [see Karlin and Taylor (1975)] 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 MED.

### 6.7 The Law of Succession

The law of succession problem is perhaps the most classical in all of probability theory [see, for example, Zabell (1989) for a lucid historical account of this rule]. 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 (41.1) 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 (1995) and Zabell (1996)] the species partition in the sample must be given by Pitman Sampling Formula (41.20). 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 MED.

### 6.8 Prime Numbers

Let \(N\) be an integer drawn at random from the set \(1, 2, \ldots, n\), and write \(N = p_1p_2p_3\cdots\), 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 \(\theta = 1\). One of the earliest investigations along these lines is Dickman (1930); see also Vershik (1986) for more recent results. Donnelly and Grimmett (1993) provide an elementary proof using size-biasing and the GEM distribution.

### 6.9 Random Mappings

The partition probability (41.4) 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 (41.4), 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 (41.19) 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 \(i \in \mathbb{Z}^+ \to \theta, i \Pr[Z_i = 1] \to \theta\) 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 MED with parameter \(\theta\) [Arratia, Barbour, and Tavaré (1995)]. Polynomial factorization satisfies \(\theta = 1\). Poisson–Dirichlet approximations for a related class of combinatorial models are given by Hansen (1994).

### BIBLIOGRAPHY


244

MULTIVARIATE EWENS DISTRIBUTION


