Submitted to Statistical Science

The ubiquitous Ewens sampling formula

Harry Crane

Rutgers, the State University of New Jersey

Abstract. Ewens's sampling formula exemplifies the harmony of mathematical theory, statistical application, and scientific discovery. The formula not only contributes to the foundations of evolutionary molecular genetics, the neutral theory of biodiversity, Bayesian nonparametrics, combinatorial stochastic processes, and inductive inference but also emerges from fundamental concepts in probability theory, algebra, and number theory. With an emphasis on its far-reaching influence throughout statistics and probability, we highlight these and many other consequences of Ewens's seminal discovery.

Key words and phrases: Ewens sampling formula, Poisson–Dirichlet distribution, random partition, coalescent process, inductive inference, exchangeability, logarithmic combinatorial structures, Chinese restaurant process, Ewens–Pitman distribution, Dirichlet process, stick breaking, permanental partition model, cyclic product distribution, clustering, Bayesian nonparametrics, α-permanent.

1. INTRODUCTION

In 1972, Warren Ewens [37] derived a remarkable formula for the sampling distribution of allele frequencies in a population undergoing neutral selection. An *allele* is a type of gene, e.g., the alleles A, B, and O in the ABO blood group, so that each gene has a particular allelic type and a sample of $n = 1, 2, \dots$ genes can be summarized by its *allelic partition* $(m_1, \dots, m_n)$, where m_1 is the number of alleles appearing exactly once, m_2 is the number of alleles appearing exactly twice, and in general m_j is the number of alleles appearing exactly j times, for each $j = 1, 2, \dots, n$. *Ewens's sampling formula* (ESF) with parameter $\theta > 0$ assigns probability

$$p(m_1, \dots, m_n; \theta) = \frac{n!}{\theta(\theta+1)\cdots(\theta+n-1)} \prod_{j=1}^{n} \frac{\theta^{m_j}}{j^{m_j} m_j!} \tag{1}$$

to each allelic partition $(m_1, \dots, m_n)$ for which $\sum_{j=1}^{n} j \cdot m_j = n$.

Derived under the assumption of selective neutrality, Equation (1) is the null hypothesis distribution needed to test the controversial neutral theory of evolution. Of the formula and its consequences, Ewens begins his abstract matter-of-factly, "In this paper a beginning is made on the sampling theory of neutral

Department of Statistics & Biostatistics 110 Frelinghuysen Road Piscataway, NJ 08854.

alleles" [37, p. 87]. Though obviously aware of its significance to the statistical theory of neutral sampling, Ewens could not have foreseen far-reaching contributions to the unified neutral theory of biodiversity [51], nonparametric Bayesian inference [2, 40], combinatorial stochastic processes [60, 73], and the philosophy of inductive inference [92], not to mention fundamental connections to the determinant function [19], Macdonald polynomials [27], and prime factorization [10, 29] in algebra and number theory. In the coming pages, we present Ewens's sampling formula in all its glory, highlighting each of these connections in further detail.

1.1 Outline

We retrace the development of Ewens's sampling formula, from neutral allele sampling and Kingman's mathematical theory of genetic diversity [60, 61, 62], to modern nonparametric Bayesian [2, 40] and frequentist [21] statistical methods, and backwards in time to the roots of probabilistic reasoning and inductive inference [8, 24, 54]. In between, Pitman's [73, 76] investigation of exchangeable random partitions and combinatorial stochastic processes unveils many more surprising connections between Ewens's sampling formula and classical stochastic process theory, while other curious appearances in algebra [19, 27] and number theory [10, 29] only add to its mystique.

1.2 Historical context

Elements of Ewens's sampling formula appeared in Yule's prior work [90] on the evolution of species, which Champernowne [16] and Simon [79] later recast in the context of income distribution in large populations and word frequencies in large pieces of text, respectively. Shortly after Ewens, Antoniak [2] independently rediscovered (1) while studying Dirichlet process priors in Bayesian statistics. We recount various other historical aspects of Ewens's sampling formula over the course of the coming pages.

1.3 Relation to prior work

Ewens & Tavaré [36, 82] have previously reviewed various structural properties of and statistical applications involving the Ewens sampling formula. The present survey provides updated and more detailed coverage of a wider range of topics, which we hope will serve as a handy reference for experts and an eye-opening introduction for beginners.

2. NEUTRAL ALLELE SAMPLING AND SYNOPSIS OF EWENS'S 1972 PAPER

2.1 The neutral Wright–Fisher evolutionary model

Population genetics theory studies the evolution of a population through changes in allele frequencies. Because many random events contribute to these changes, the theory relies on stochastic models. The simplest of these is the Wright–Fisher model, following its independent introduction by Fisher [41] and Wright [89].

The Wright–Fisher model concerns a diploid population, i.e., a bi-parental population in which every individual has two genes, one derived from each parent, at any gene locus. The population is assumed to be of fixed size N, so that there are $2N$ genes at each gene locus in any generation. The generic Wright–Fisher model assumes that the $2N$ genes in any offspring generation are obtained by sampling uniformly with replacement from the $2N$ genes in the parental generation.

A gene comprises a sequence of ℓ DNA nucleotides, where ℓ is typically on the order of several thousand, and thus each gene has exactly one of the possible 4^ℓ allelic types. The large number of allelic types motivates the *infinitely many alleles* assumption, by which each transmitted gene is of the same type as its parental gene with probability $1-u$ and mutates independently with probability u to a new allelic type "not currently existing (nor previously existing) in the population" [37, p. 88].

Under these assumptions, Ewens [37] derives Equation (1) by a partly heuristic argument made precise by Karlin and McGregor [55]. The parameter θ in Ewens's sampling formula is equal to $4Nu$ and, therefore, admits an interpretation in terms of the mutation rate. In more general applications, θ is best regarded as an arbitrary parameter.

Ewens goes on to discuss both deductive and inductive questions about the sample and population and, in the latter half of [37], addresses issues surrounding statistical inference and hypothesis testing in population genetics. Among all these contributions, Ewens's discussion in the opening pages about the mean number of alleles and the distribution of allele frequencies has had the most lasting impact.

2.2 The mean number of alleles

Equation (1) leads directly to the probability distribution of the number of different alleles K in the sample as

$$\mathbb{P}\{K = k\} = |S_n^k| \frac{\theta^k}{\theta(\theta+1)\cdots(\theta+n-1)},$$

where S_n^k is the (n,k)-Stirling number of the first kind [80]:A008275. Together with Equation (1), the above expression implies that K is a sufficient statistic for θ, a point discussed later, and also leads to the expression

$$\frac{\theta}{\theta} + \frac{\theta}{\theta+1} + \cdots + \frac{\theta}{\theta+n-1} \sim \theta \log(n)$$

for the mean of K, where $a \sim b$ indicates that $a/b \to 1$ as $n \to \infty$. Contrasting this with the mean number of distinct alleles in a population of size N,

$$\frac{\theta}{\theta} + \frac{\theta}{\theta+1} + \cdots + \frac{\theta}{\theta+2N-1} \sim \theta \log(2N),$$

Ewens approximates the mean number of different alleles in the population that do not appear in the sample by $\theta \log(2N/n)$.

2.3 Predictive probabilities

Invoking "a variant of the 'coupon collector's problem' (or the 'law of succession')" [37, p. 94], Ewens calculates the probability that "the $(j+1)$th gene drawn is of an allelic type not observed on the first j draws" as $\theta/(\theta+j)$. Conversely, the probability that the $(j+1)$st gene is of a previously observed allelic type is

$j/(\theta + j)$. These predictive probabilities precede Dubins and Pitman's Chinese restaurant seating rule (Section 4) and posterior probabilities based on Ferguson's Dirichlet process prior (Section 6) and are also closely associated with De Morgan's rule of succession for inductive questions (Section 7).

2.4 The coalescent

More than describing the allelic configuration of genes at any given time, the neutral Wright–Fisher model describes the evolution of a population of size N with nonoverlapping generations. Under this evolution, Equation (1) acts as the stationary distribution for a sample of n genes from a large population. In an effort to better understand how (1) arises from these dynamics, Griffiths [46] and Kingman [63] considered the behavior of allelic frequencies under the infinite population diffusion approximation. Perhaps the major advance in evolutionary population genetics in the last three decades, Kingman's coalescent has resulted in the widely adopted coalescent theory within population genetics [86] as well as the mathematical study of partition-valued and more general combinatorial stochastic processes [9, 76].

Under the dynamics of Section 2.1, each of the $2N$ genes in the current generation has a parent gene in the previous generation. Tracing these parental relationships backwards in time produces the ancestral lineages of the current $2N$ genes. Thus, the number of children genes X of a typical gene is a Binomial random variable with success probability $1/(2N)$, i.e.,

$$\mathbb{P}\{X = k\} = \binom{2N}{k}(2N)^{-k}(1-(2N)^{-1})^{N-k}, \quad k = 0, 1, \ldots, 2N, \tag{2}$$

and so the probability that two genes in the current generation have the same parent is $1/(2N)$. More generally, the number of generations Y for which the ancestral lines of two genes have been distinct follows the Geometric distribution with success probability $1/(2N)$, i.e.,

$$\mathbb{P}\{Y \geq k\} = (1-(2N)^{-1})^k, \quad k = 0, 1, \ldots. \tag{3}$$

From (2) and (3), the probability that the ancestral lines of ℓ genes have been distinct for k generations is

$$(1-1/(2N))^k(1-2/(2N))^k \cdots (1-(\ell-1)/(2N))^k.$$

The coalescent process arises as a natural infinite population diffusion approximation to the Wright–Fisher model by taking $N \to \infty$ and putting $k = \lfloor 2Nt \rfloor$ for $t \geq 0$. Under this regime, we obtain the limiting probability that the lineages of ℓ genes remain distinct for $t \geq 0$ time units as

$$\lim_{N\to\infty, k/(2N)\to t} \prod_{j=1}^{\ell-1}(1-j/(2N))^k = \exp\{-t\ell(\ell-1)/2\}.$$

Realizing that this equals the distribution of the minimum of $\binom{\ell}{2}$ independent standard Exponential random variables, Kingman [63] arrived at his description of *the coalescent*, according to which distinct lineages merge independently at the times of standard Exponential random variables.

Although we have focused mainly on its original derivation in the context of the Wright–Fisher model, Ewens's sampling formula applies for a wide range of neutral models [59]. In their treatment of Macdonald polynomials (Section 8.3), Diaconis & Ram [27] attribute these "myriad practical appearances [to] its connection with Kingman's coalescent process[...]." Ethier & Griffiths's [32] formula for the transition function of the Fleming–Viot process [43] provides yet another link between the coalescent, Dirichlet processes, the Poisson–Dirichlet distribution, and Ewens's sampling formula.

2.5 Legacy in theoretical population genetics

The sufficiency of K for θ led Ewens [37] and Watterson [87, 88] to objective tests of the controversial neutral theory of evolution [58]. Plainly, sufficiency implies that the conditional distribution of $(m_1, \dots, m_n)$ given K is independent of θ, so that the unknown parameter θ need not be estimated and is not involved in any test based on the conditional distribution of $(m_1, \dots, m_n)$ given K. Beyond the realm of Ewens's seminal work on testing, Christiansen [17] states that Ewens [37] "laid the foundations for modern molecular population genetics." Nevertheless, in the early 1980's Ewens shifted his focus to mapping genes associated to human diseases. He is partly responsible for the transmission-disequilibrium test [81], which has been used to locate at least fifty such genes. The paper [81] that introduced the transmission-disequilibrium test was chosen as one of the top ten classic papers in the *American Journal of Human Genetics*, 1949–2014.

In general biology, Ewens's paper marks a seminal contribution to Hubbell's neutral theory of biodiversity [51], and the foundations it laid have been refined by novel sampling formulas [33, 35] and statistical tests for neutrality [34] in ecology. For the rest of the paper, we focus on applications of Equation (1) in other areas, mentioning other biological applications only briefly.

3. CHARACTERISTIC PROPERTIES OF EWENS'S SAMPLING FORMULA

3.1 Partition structures

From an allelic partition $(m_1, \dots, m_n)$ of $n \geq 1$, we obtain an allelic partition $(m'_1, \dots, m'_{n-1})$ of $n-1$ by choosing J randomly with probability

$$\mathbb{P}\{J = j \mid (m_1, \dots, m_n)\} = j \cdot m_j / n, \quad j = 1, \dots, n, \tag{4}$$

and putting

$$m'_j = \begin{cases} m_j - 1, & J = j \\ m_j + 1, & J = j+1 \\ m_j, & \text{otherwise.} \end{cases} \tag{5}$$

Alternatively, (5) is the allelic partition obtained by choosing a gene uniformly at random and removing it from the sample.

Theorem 3.1 (Kingman [60, 61]). *Let $(m_1, \dots, m_n)$ be a random allelic partition from Ewens's sampling formula (1) with parameter $\theta > 0$. Then $(m'_1, \dots, m'_{n-1})$ obtained as in (5) is also distributed according to Ewens's sampling formula with parameter $\theta > 0$.*

We call a family of distributions $(p_n)_{n\geq 1}$ a *partition structure* if $(m_1, \dots, m_n) \sim p_n$ implies $(m'_1, \dots, m'_{n-1}) \sim p_{n-1}$ for all $n \geq 2$. This definition implies that $(p_n)_{n\geq 1}$ is consistent under uniform deletion of any number of genes and, thus, agrees with Kingman's original definition [60], which generalizes the outcome in Theorem 3.1.

Kingman's study of partition structures anticipates his *paintbox process correspondence*, by which he proves a de Finetti-type representation for all infinite exchangeable random set partitions. Kingman's correspondence establishes a link between Ewens's sampling formula and the Poisson–Dirichlet distribution, which in turn broadens the scope of Equation (1); see Section 4.1 below.

3.2 Random set partitions

In many respects, partition structures are more naturally developed as distributions on partitions of the set $[n] = \{1, \dots, n\}$. Instead of summarizing the sample of genes by the allelic partition $(m_1, \dots, m_n)$, we can label genes distinctly $1, \dots, n$ and assign each an allelic type. At each generation, the genes segregate into subsets $B_1, \dots, B_k$, called *blocks*, such that i and j in the same block indicates that genes i and j have the same allelic type. The resulting collection $\pi = \{B_1, \dots, B_k\}$ of non-empty, disjoint subsets with $B_1 \cup \cdots \cup B_k = [n]$ is a *(set) partition* of $[n]$.

The ordering of $B_1, \dots, B_k$ in π is inconsequential, so we follow convention and list blocks in ascending order of their smallest element. For example, there are five partitions of the set $\{1, 2, 3\}$,

$$\{\{1,2,3\}\}, \quad \{\{1\},\{2,3\}\}, \quad \{\{1,2\},\{3\}\}, \quad \{\{1,3\},\{2\}\}, \quad \{\{1\},\{2\},\{3\}\},$$

but only three allelic partitons of size 3,

$$(3,0,0), \quad (1,1,0), \quad (0,0,1).$$

Each set partition $\pi = \{B_1, \dots, B_k\}$ corresponds uniquely to an allelic partition $\mathbf{n}(\pi) = (m_1, \dots, m_n)$, where m_j counts the number of blocks of size j in π, e.g.,

$$\begin{aligned} \mathbf{n}(\{\{1,2,3\}\}) &= (0,0,1), \\ \mathbf{n}(\{\{1,2\},\{3\}\}) = \mathbf{n}(\{\{1\},\{2,3\}\}) = \mathbf{n}(\{\{1,3\},\{2\}\}) &= (1,1,0), \quad \text{and} \\ \mathbf{n}(\{\{1\},\{2\},\{3\}\}) &= (3,0,0). \end{aligned}$$

Conversely, every allelic partition $(m_1, \dots, m_n)$ corresponds to the set of all partitions π for which $\mathbf{n}(\pi) = (m_1, \dots, m_n)$.

Because every individual draws its parent genes independently and uniformly in the Wright–Fisher model, set partitions corresponding to the same allelic partition occur with the same probability. Consequently, we can generate a random set partition Π_n by first drawing an allelic partition $(m_1, \dots, m_n)$ from Ewens's sampling formula (1) and then selecting uniformly among partitions π for which $\mathbf{n}(\pi) = (m_1, \dots, m_n)$. The resulting random set partition Π_n follows the so-called *Ewens distribution* with parameter $\theta > 0$,

$$\mathbb{P}\{\Pi_n = \{B_1, \dots, B_k\}\} = \frac{\theta^k}{\theta(\theta+1)\cdots(\theta+n-1)} \prod_{j=1}^{k} (\#B_j - 1)!, \tag{6}$$

where $\#B_j$ is the cardinality of block B_j for each $j = 1, \dots, k$.

Simple enumeration and the law of total probability connects (1) and (6). As an exercise, the reader can verify that each allelic partition $(m_1, \ldots, m_n)$ corresponds to $n!/\prod_{j=1}^{n} j!^{m_j} m_j!$ set partitions through $\mathbf{n}$ and, therefore, the conditional distribution of a partition drawn uniformly among all π whose block sizes form allelic partition $(m_1, \ldots, m_n)$ is

$$\mathbb{P}\{\Pi_n = \pi \mid \mathbf{n}(\Pi_n) = (m_1, \ldots, m_n)\} = \frac{1}{n!}\prod_{j=1}^{n} j!^{m_j} m_j!, \quad \mathbf{n}(\pi) = (m_1, \ldots, m_n).$$

3.3 Exchangeability

The distribution in (6) depends only on the allelic partition induced by the block sizes of Π_n. Therefore, for any permutation $\sigma : [n] \to [n]$, the *relabeling* Π_n^σ obtained by first taking Π_n distributed as in (6) and then putting $\sigma(i)$ and $\sigma(j)$ in the same block of Π_n^σ if and only if i and j are in the same block of Π_n is also distributed as in (6). As is natural in the genetics setting, the labels $1, \ldots, n$ distinguish between genes but otherwise can be assigned arbitrarily.

More generally, a random partition Π_n is *exchangeable* if its distribution is invariant under relabeling by any permutation $\sigma : [n] \to [n]$, that is,

$$\mathbb{P}\{\Pi_n = \pi\} = \mathbb{P}\{\Pi_n = \pi^\sigma\} \quad \text{for all permutations } \sigma : [n] \to [n].$$

Pitman's *exchangeable partition probability function* (EPPF) [73] captures the notion of exchangeability through a function $P_n(m_1, \ldots, m_n)$ on allelic partitions. In particular, Π_n is exchangeable if and only if there exists an EPPF P_n such that

$$\mathbb{P}\{\Pi_n = \pi\} = P_n(\mathbf{n}(\pi)) \quad \text{for all partitions of } [n].$$

Exchangeability of the Ewens distribution is clear from the closed form expression in (6), which depends on Π_n only through its block sizes. In general, every probability distribution $p_n(\cdot)$ on allelic partitions of n determines an EPPF P_n by

$$(7) \qquad P_n(\mathbf{n}(\pi)) = p_n(m_1, \ldots, m_n) \times \frac{1}{n!}\prod_{j=1}^{n} j!^{m_j} m_j!, \quad \mathbf{n}(\pi) = (m_1, \ldots, m_n).$$

3.4 Consistency under subsampling

Above all, Kingman's definition of a partition structure emphasizes the "need for consistency between different sample sizes" [61, p. 374]. By subsampling $[m] \subset [n]$, a partition $\pi = \{B_1, \ldots, B_k\}$ of $[n]$ *restricts* to a partition of $[m]$ by

$$\pi_{|[m]} = \{B_1 \cap [m], \ldots, B_k \cap [m]\} \setminus \{\emptyset\}.$$

For example, the restrictions of $\pi = \{\{1,4,7,8\},\{2,3,5\},\{6\}\}$ to samples of size $m = 7, 6, 5$, respectively, are

$$\pi_{|[7]} = \{\{1,4,7\},\{2,3,5\},\{6\}\}, \quad \pi_{|[6]} = \{\{1,4\},\{2,3,5\},\{6\}\}, \quad \pi_{|[5]} = \{\{1,4\},\{2,3,5\}\}.$$

To satisfy the partition structure requirement, the sampling distribution of Π_m must coincide with the marginal distribution of $\Pi_{n|[m]}$, the partition induced by subsampling m genes from a sample of size $n \geq m$. A family of random set partitions $(\Pi_n)_{n\geq 1}$ is *consistent under subsampling*, or *sampling consistent*, if $\Pi_{n|[m]}$ is distributed the same as Π_m for all $n \geq m \geq 1$. Similarly, a family of distributions is sampling consistent if it governs a consistent family of random partitions. Any partition structure $(p_n)_{n\geq 1}$ determines a family of exchangeable, consistent EPPFs through (7). In particular, Ewens's sampling formula (1) determines a partition structure and the Ewens distribution in (6) is derived from (1) via (7); thus, the family of Ewens distributions is consistent under subsampling.

Sampling consistency elicits an interpretation of the distributions of $(\Pi_n)_{n\geq 1}$ as the sampling distributions induced by a data generating process for the whole population. Inductively, the sampling distributions of $(\Pi_n)_{n\geq 1}$ permit a sequential construction: given $\Pi_n = \pi$, we generate Π_{n+1} from the conditional distribution among all partitions of $[n+1]$ that restrict to π under subsampling, i.e.,

$$\mathbb{P}\{\Pi_{n+1} = \pi' \mid \Pi_n = \pi\} = \begin{cases} \mathbb{P}\{\Pi_{n+1} = \pi'\}/\mathbb{P}\{\Pi_n = \pi\}, & \pi'_{|[n]} = \pi, \\ 0, & \text{otherwise.} \end{cases}$$

Thus, a consistent family of partitions implies the existence of conditional distributions for predictive inference and the sequence $(\Pi_n)_{n\geq 1}$ determines a unique *infinite random partition* Π_∞ of the positive integers $\mathbb{N} = \{1, 2, \ldots\}$. In the special case of Ewens's distribution, these predictive probabilities determine the Chinese restaurant process (Section 4.5), which figures into nonparametric Bayesian posterior inference via the Dirichlet process prior (Section 6.1).

3.5 Self-similarity

In the Wright–Fisher model, sub-populations exhibit the same behavior as the population-at-large and different genes do not interfere with each other. Together, these comprise the statistical property of *self-similarity*.

Formally, set partitions are partially ordered by the *refinement* relation: we write $\pi \leq \pi'$ if every block of π is a subset of some block of π'. For example, $\pi = \{\{1,2\},\{3,4\},\{5\}\}$ refines $\pi' = \{\{1,2,5\},\{3,4\}\}$ but not $\pi'' = \{\{1,3\},\{2,4,5\}\}$, because $\{1,2\}$ is not a subset of $\{1,3\}$ or $\{2,4,5\}$. Let $P(\cdot)$ be an EPPF for an infinite exchangeable random partition, so that P determines an exchangeable probability distribution on partitions of any finite size by $\mathbb{P}\{\Pi_n = \pi\} = P(\mathbf{n}(\pi))$. The family of random set partitions $(\Pi_n)_{n\geq 1}$ is *self-similar* if for all $n = 1, 2, \ldots$ and all set partitions π of $[n]$

$$\mathbb{P}\{\Pi_n = \pi \mid \Pi_n \leq \pi'\} = \prod_{b\in\pi'} P(\mathbf{n}(\pi_{|b})), \quad \pi \leq \pi'. \tag{8}$$

In other words, given that Π_n is a refinement of π, the further breakdown of elements within each block of π occurs independently of other blocks and with the same distribution determined by P. The reader can verify that (6) satisfies condition (8).

3.6 Non-interference

A longstanding question in ecology concerns interactions between different species. In this context, we regard the elements $1, 2, \ldots$ as labels for different specimens, instead of genes, so that the allelic partition $(m_1, \ldots, m_n)$ counts the number of species that appear once, twice, and so on in a sample. Given an arbitrary partition structure $(p_n)_{n\geq 1}$ and $(m_1, \ldots, m_n) \sim p_n$, we choose an index

$J = r$ as in (4) and, instead of reducing to an allelic partition of $n-1$ as in (5), we define $m^* = (m^*_1, \ldots, m^*_{n-r})$ by

$$m^*_j = \begin{cases} m_j - 1, & J = j, \\ m_j, & \text{otherwise}, \end{cases}$$

to obtain an allelic partition of $n - r$. In effect, we remove all specimens with the same species as one chosen uniformly at random among $1, \ldots, n$. If for every $n = 1, 2, \ldots$ the conditional distribution of m^* given $J = r$ is distributed according to p_{n-r}, the family $(p_n)_{n\geq 1}$ satisfies *non-interference*. Kingman [60] showed that Ewens's sampling formula (1) is the only partition structure with the non-interference property.

3.7 Exponential families, Gibbs partitions, and product partition models

The family of Ewens distributions on set partitions can be expressed as an exponential family with natural parameter $\log(\theta)$ and canonical sufficient statistic the number of blocks of Π_n:

$$(9) \qquad \mathbb{P}\{\Pi_n = \pi\} = \exp\{\#\pi \log\theta - \sum_{j=0}^{n-1} \log(\theta + j)\} \times \prod_{b\in\pi} (\#b - 1)!,$$

where $\#\pi$ denotes the number of blocks of π. Within statistical physics, (9) can be re-written in *canonical Gibbs form* as

$$(10) \qquad \mathbb{P}\{\Pi_n = \pi\} = Z_n^{-1} \prod_{b\in\pi} \psi(\#b),$$

for non-negative constants $\psi(k) = \theta \cdot (k-1)!$ and normalizing constant $Z_n = \theta(\theta+1)\cdots(\theta+n-1)$. The following theorem distinguishes the Ewens family among this class of Gibbs distributions.

THEOREM 3.2 (Kerov [56]). *A family $(P_n)_{n\geq 1}$ of Gibbs distributions (10) with common weight sequence $\{\psi(k)\}_{k\geq 1}$ is exchangeable and consistent under subsampling if and only if there exists $\theta > 0$ such that $\psi(k) = \theta \cdot (k-1)!$ for all $k \geq 1$.*

Without regard for statistical or physical properties of (10), Hartigan [47] proposed the class of *product partition models* for certain statistical clustering applications. For a collection of cohesion functions $c(b)$, $b \subseteq [n]$, the product partition model assigns probability

$$\mathbb{P}\{\Pi_n = \pi\} \propto \prod_{b\in\pi} c(b)$$

to each partition π of $[n]$. Clearly, the product partition model is exchangeable only if $c(b)$ depends only on the cardinality of $b \subseteq [n]$. Kerov's work (Theorem 3.2) establishes that the only non-degenerate, exchangeable, consistent product partition model is the family of Ewens distributions in (6).

3.8 Logarithmic combinatorial structures

For $\theta > 0$, let $Y_1, Y_2, \ldots$ be independent random variables for which Y_j has the Poisson distribution with parameter θ/j for each $j = 1, 2, \ldots$. Given $\sum_{j=1}^n j \cdot Y_j = n$, $(Y_1, \ldots, Y_n)$ determines a random allelic partition of n with distribution

$$\mathbb{P}\left\{(Y_1, \ldots, Y_n) = (m_1, \ldots, m_n) \,\middle|\, \sum_{j=1}^n j \cdot Y_j = n\right\} \propto \prod_{j=1}^n \frac{\theta^{m_j}}{j^{m_j} m_j!} e^{-\theta/j} \propto \prod_{j=1}^n \frac{\theta^{m_j}}{j^{m_j} m_j!}, \tag{11}$$

that is, Ewens's sampling formula with parameter $\theta > 0$.

From the above description, the collection $(\Pi_n)_{n\geq 1}$ of Ewens partitions is a special case of a logarithmic combinatorial structure, which Arratia, Barbour & Tavaré [4] define for set partitions as follows. Let $(\Pi_n)_{n\geq 1}$ be a collection of random set partitions and, for each $n \geq 1$, let $(N_{n,1}, \ldots, N_{n,n})$ be the random allelic partition determined by the block sizes of Π_n. Then $(\Pi_n)_{n\geq 1}$ is a *logarithmic combinatorial structure* if for every $n = 1, 2, \ldots$ the allelic partition $(N_{n,1}, \ldots, N_{n,n})$ satisfies the *conditioning relation*

$$\mathbb{P}\{N_{n,1} = m_1, \ldots, N_{n,n} = m_n\} = \mathbb{P}\{Y_1 = m_1, \ldots, Y_n = m_n \mid \sum_{j=1}^n j \cdot Y_j = n\}$$

for some sequence $Y_1, Y_2, \ldots$ of independent random variables on $\{0, 1, \ldots\}$ that satisfies the *logarithmic condition*

$$\lim_{n\to\infty} n\mathbb{P}\{Y_n = 1\} = \lim_{n\to\infty} n\mathbb{E}Y_n > 0.$$

Both conditions are plainly satisfied by the independent Poisson sequence above.

Arratia, Barbour & Tavaré [3] further established the following stronger result according to which the block sizes of a random Ewens partition can be well approximated by independent Poisson random variables as the sample size grows.

Theorem 3.3 (Arratia, Barbour & Tavaré [3]). *For $n \geq 1$, let $N_{n,j}$ be the number of blocks of size j in a Ewens(θ) partition of $[n]$. Then $(N_{n,j})_{j\geq 1} \to_{\mathcal{D}} (Y_1, Y_2, \ldots)$ as $n \to \infty$, where $Y_1, Y_2, \ldots$ are independent Poisson random variables with $E(Y_j) = \theta/j$ and $\to_{\mathcal{D}}$ denotes* convergence in distribution*.*

In the above theorem, $N_{n,j}$ counts the number of blocks of size j in a random partition of $\{1, \ldots, n\}$. More generally, $N_{n,j}$ may count the number of components of size j in an arbitrary structure of size n, e.g., the components of size j in a random graph or a random mapping. Arratia, Barbour & Tavaré's monograph [5] relates the component sizes of various structures to Ewens's sampling formula, e.g., 2-regular graphs (with $\theta = 1/2$) and properties of monic polynomials (with $\theta = 1$). Aldous [1, Lemma 11.23] previously showed that the component sizes of the directed graph induced by a uniform random mapping $[n] \to [n]$ converge in distribution to the component sizes from Ewens's sampling formula with $\theta = 1/2$.

4. SEQUENTIAL CONSTRUCTIONS AND URN SCHEMES

4.1 Poisson–Dirichlet distribution

Let

$$\mathcal{S}^{\downarrow} = \left\{(s_1, s_2, \ldots) : s_1 \geq s_2 \geq \cdots \geq 0, \sum_{k \geq 1} s_k \leq 1\right\}$$

be the ranked-simplex and, for $s \in \mathcal{S}^{\downarrow}$, write $\Pi_\infty \sim \varrho_s$ to signify an infinite random partition generated as follows. Let $X_1, X_2, \ldots$ be independent random variables with distributions

$$\mathbb{P}\{X_i = j \mid s\} = \begin{cases} s_j, & j \geq 1, \\ 1 - \sum_{k \geq 1} s_k, & j = -i, \\ 0, & \text{otherwise}. \end{cases} \tag{12}$$

From $(X_1, X_2, \ldots)$, we define the *s-paintbox* $\Pi_\infty \sim \varrho_s$ by putting

$$i \text{ and } j \text{ in the same block of } \Pi_\infty \quad \text{if and only if} \quad X_i = X_j. \tag{13}$$

Notice that $s_0 = 1 - \sum_{k \geq 1} s_k$ is the probability that $X_i = -i$ for each $i = 1, 2, \ldots$ and, therefore, corresponds to the probability that element i appears as a singleton in Π_∞. By the law of large numbers and Kingman's correspondence, each block of an infinite exchangeable partition is either a singleton or is infinite; there can be no blocks of size two, three, etc., or with zero limiting frequency.

THEOREM 4.1 (Kingman's correspondence [61]). *Let Π_∞ be an infinite exchangeable partition. Then there exists a unique probability measure ν on $\mathcal{S}^{\downarrow}$ such that $\Pi_\infty \sim \varrho_\nu$, where*

$$\varrho_\nu(\cdot) = \int_{\mathcal{S}^{\downarrow}} \varrho_s(\cdot) \nu(ds) \tag{14}$$

is the mixture of s-paintbox measures with respect to ν.

By Kingman's correspondence, every exchangeable random partition of $\mathbb{N}$ can be constructed by first sampling $S \sim \nu$ and then "painting" elements $1, 2, \ldots$ according to (12). In the special case of Ewens's sampling formula, the mixing measure ν is called the *Poisson–Dirichlet distribution* with parameter $\theta > 0$.

Kingman further showed that if a sequence of populations of growing size is such that for each population the allelic partition from a sample of n genes obeys Ewens's sampling formula, then the limiting distribution of allele frequencies is the Poisson–Dirichlet distribution with parameter θ. In genetics, the Poisson–Dirichlet distribution can be viewed as an infinite population result for selectively neutral alleles in the infinitely many alleles setting. Within statistics, the Poisson–Dirichlet distribution is the prior distribution over the set of all paintboxes whose colors occur according to the proportions of $s \in \mathcal{S}^{\downarrow}$. Below, we lay bare several instances of the Poisson–Dirichlet distribution throughout mathematics. The limit of the Dirichlet-Multinomial process offers perhaps the most tangible interpretation of the Poisson–Dirichlet distribution.

4.2 Dirichlet-Multinomial process

For $\alpha_1, \ldots, \alpha_k > 0$, the $(k-1)$-*dimensional Dirichlet distribution with parameter* $(\alpha_1, \ldots, \alpha_k)$ has density

$$(15) \quad f(s_1, \ldots, s_k; \alpha_1, \ldots, \alpha_k) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_k)}{\prod_{i=1}^k \Gamma(\alpha_i)} s_1^{\alpha_1 - 1} \cdots s_k^{\alpha_k - 1}, \qquad \begin{array}{c} s_1 + \cdots + s_k = 1 \\ s_1, \ldots, s_k \geq 0, \end{array}$$

where $\Gamma(s) = \int_0^\infty x^{s-1} e^{-x} dx$ is the gamma function. Given $S = (S_1, \ldots, S_k)$ from the above density, we draw $X_1, X_2, \ldots$ conditionally independently from

$$\mathbb{P}\{X_1 = j \mid S = (s_1, \ldots, s_k)\} = s_j, \quad j = 1, \ldots, k,$$

and define a random partition Π_∞ as in (13). With $\alpha < 0$, $\alpha_1 = \cdots = \alpha_k = -\alpha$, and $n_j = \sum_{i=1}^n \mathbf{1}\{X_i = j\}$, the count vector $(n_1, \ldots, n_k)$ for sample of size n has unconditional probability

$$\int_{[0,1]^k} \frac{\Gamma(-k\alpha)}{\Gamma(-\alpha)^k} s_1^{-\alpha + n_1 - 1} \cdots s_k^{-\alpha + n_k - 1} ds_1 \cdots ds_k = \frac{1}{(-k\alpha)^{\uparrow n}} \prod_{j=1}^k (-\alpha)^{\uparrow n_j},$$

where $\alpha^{\uparrow j} = \alpha(\alpha+1) \cdots (\alpha + j - 1)$ is the rising factorial function. Here we specify α to be negative in order to comply with the parameterization of the forthcoming Ewens–Pitman distribution (Section 5.1)

In determining a random partition Π_n based on $X_1, \ldots, X_n$, we disregard the specific values of $X_1, \ldots, X_n$ and only retain the equivalence classes. If j distinct values appear among $X_1, \ldots, X_n$, there are $k^{\downarrow j} = k(k-1) \cdots (k - j + 1)$ possible assignments that induce the same partition of $[n]$; whence,

$$(16) \qquad \mathbb{P}\{\Pi_n = \pi\} = \frac{k^{\downarrow \#\pi}}{(-k\alpha)^{\uparrow n}} \prod_{b \in \pi} (-\alpha)^{\uparrow \#b} = \frac{(-k\alpha/\alpha)^{\uparrow \#\pi}}{(-k\alpha)^{\uparrow n}} \prod_{b \in \pi} -(-\alpha)^{\uparrow \#b}.$$

The Poisson–Dirichlet(θ) distribution corresponds to Ewens's one parameter family (6) and can be viewed as a limiting case of the above Dirichlet-Multinomial construction in the following sense. Let $\theta > 0$ and, for each $m = 1, 2, \ldots$, let $\Pi_{n,m}$ have the Dirichlet-Multinomial distribution in (16) with parameters $\alpha = -\theta/m$ and $k = m$. The distributions of $\Pi_{n,m}$ satisfy

$$\mathbb{P}\{\Pi_{n,m} = \pi\} = \frac{m^{\downarrow \#\pi}}{\theta^{\uparrow n}} \prod_{b \in \pi} (\theta/m)^{\uparrow \#b}$$

and, therefore, $\Pi_{n,m}$ converges in distribution to a Ewens(θ) partition of $[n]$ as $m \to \infty$.

In Kingman's paintbox process, $\Pi_{n,m}$ is the mixture with respect to the distribution of decreasing order statistics of the $(m-1)$-dimensional Dirichlet distribution with parameter $(\theta/m, \ldots, \theta/m)$, whereas a Ewens($\theta$) partition is the mixture with respect to the Poisson–Dirichlet(θ) distribution. By the bounded convergence theorem, Poisson–Dirichlet(θ) is the limiting distribution of the decreasing order statistics of $(m-1)$-dimensional Dirichlet$(\theta/m, \ldots, \theta/m)$ distributions as $m \to \infty$. This connection between Ewens's sampling formula and the Dirichlet-Multinomial construction partially explains the utility of the Chinese restaurant process in Bayesian inference (Section 6.1). Consult Feng [39] for more details on the Poisson–Dirichlet distribution and its connections to diffusion processes.

4.3 Hoppe's urn

In the paintbox process, we construct a random partition by sampling $X_1, X_2, \ldots$ conditionally independently and defining its blocks as in (13). Alternatively, Hoppe [49] devised a Pólya-type urn scheme by which (6) arises by sampling with reinforcement from an urn.

Initiate an urn with a single black ball with label 0 and weight $\theta > 0$. Sequentially for each $n = 1, 2, \ldots$, choose a ball with probability proportional to its weight and replace it along with a new ball labeled n, weighted 1, and colored

- the same as the chosen ball, if not black, or
- differently from all other balls in the urn, if the chosen ball is black.

The above scheme determines a partition of $\{1, 2, \ldots\}$ for which i and j are in the same block if and only if the balls labeled i and j have the same color. Hoppe showed that the color composition $(m_1, \ldots, m_n)$ is distributed as in Equation (1), where m_j is the number of colors represented by exactly j of the first n non-black balls.

From this point on, we leave behind the interpretation of $1, 2, \ldots$ as labels for genes, as in Ewens's original context, in favor of a more generic setting in which $1, 2, \ldots$ are themselves abstract items or elements, e.g., labels of balls in Hoppe's urn.

4.4 De Morgan's process

Some 150 years before Hoppe, De Morgan [24] posited a similar sequential scheme for explaining how to update conditional probabilities for events that have not yet occurred and are not even known to exist:

> "When it is known beforehand that either A or B *must* happen, and out of $m + n$ times A has happened m times, and B n times, then[...]it is $m + 1$ to $n + 1$ that A will happen the next time. But suppose we have no reason, except that we gather from the observed event, to know that A or B must happen; that is, suppose C or D, or E, etc. might have happened: then the next event may be either A or B, or a new species, of which it can be found that the respective probabilities are proportional to $m + 1$, $n + 1$, and 1[...]." (De Morgan, [24, p. 66])

Thus, De Morgan considers situations for which we do not even know all the possible outcomes in advance, as opposed to binary outcomes such as whether the sun will rise or not or whether a coin will land on heads or tails. On the $(n + 1)$st trial, De Morgan assigns probability $1/(n + t + 1)$ to the event that a new type is observed and $n_j/(n + t + 1)$ to the event that a type with n_j prior occurrences is observed. With $\theta = 1$ and $t = 0$, De Morgan's and Hoppe's update probabilities coincide. This framework is sometimes called the *species sampling problem*—before encountering an animal of a new species we are not aware that the species exists—and requires the use of exchangeable random partitions instead of exchangeable random variables [92].

In deriving (1), Ewens invokes "a variant of the 'coupon collector's problem' (or the 'law of succession')" [37, p. 94] and must have been aware of the sequential construction of (1) via Hoppe's urn or, equivalently, the Chinese restaurant process from the coming section. In fact, the update probabilities of Ewens's sampling formula are distinguished within the broader context of rules of succession: if the conditional probability that item $n + 1$ is of a new type depends only on n, then it must have the form $\theta/(\theta + n)$ for some $\theta > 0$ [28]; moreover, if the conditional probability that the $(n+1)$st item is of a type seen m times previously depends only on m and n, then the underlying sampling distribution must be Ewens's sampling formula. If, in addition to depending on m, the conditional probability at stage $n+1$ also depends on the number of species observed so far, then the underlying distribution has the two-parameter Ewens–Pitman distribution, which we discuss throughout Section 5. This latter point relies on Johnson's [54] sufficientness postulate, which we discuss further in Section 7.

4.5 Chinese restaurant process

Dubins and Pitman, see e.g. [1, Section 11.19], proposed the *Chinese restaurant process*, a sampling scheme equivalent to Hoppe's urn above. Imagine a restaurant in which customers are labeled according to the order in which they arrive: the first customer is labeled 1, the second is labeled 2, and so on. If the first n customers are seated at $m \geq 1$ different tables, the $(n+1)$st customer sits

- at a table occupied by $t \geq 1$ customers with probability $t/(\theta + n)$ or
- at an unoccupied table with probability $\theta/(\theta + n)$,

where $\theta > 0$. By regarding the balls in Hoppe's urn as customers in a restaurant, it is clear that the Chinese restaurant process and Hoppe's urn scheme are identical and, thus, the allelic partition $(m_1, \ldots, m_n)$, for which m_j counts the number of tables with j customers, is distributed according to Equation (1).

5. THE TWO-PARAMETER EWENS–PITMAN DISTRIBUTION

5.1 Ewens–Pitman two-parameter family

In some precise sense, the distribution in (16) can be viewed as a specialization of (6) to the case of a partition with a bounded number of blocks. Both (6) and (16) are special cases of Pitman's two-parameter extension to Ewens's sampling formula; see [76] for an extensive list of references.

With (α, θ) satisfying either

- $\alpha < 0$ and $\theta = -k\alpha$, for some $k = 1, 2, \ldots$, or
- $0 \leq \alpha \leq 1$ and $\theta > -\alpha$,

the *Ewens–Pitman distribution* with parameter (α, θ) assigns probability

$$\mathbb{P}\{\Pi_n = \pi\} = \frac{(\theta/\alpha)^{\uparrow \#\pi}}{\theta^{\uparrow n}} \prod_{b \in \pi} -(-\alpha)^{\uparrow \#b}. \tag{17}$$

When $\alpha = 0$, (17) coincides with Equation (6); and when $\alpha < 0$ and $\theta = -k\alpha$, (17) simplifies to (16). In terms of the paintbox process (14), the mixing measure of an infinite partition drawn from (17) is the *(two-parameter) Poisson–Dirichlet distribution* with parameter (α, θ) [72, 77]. The one-parameter Poisson–Dirichlet(θ) distribution mentioned previously is the specialization of the two-parameter case with $\alpha = 0$ and may also be called the Poisson–Dirichlet$(0, \theta)$ distribution.

5.2 Role of parameters

A more general Chinese restaurant-type construction, as in Section 4.5, elucidates the meaning of parameters α and θ in (17). If the first n customers are seated at $m \geq 1$ different tables, the $(n+1)$st customer sits

- at a table occupied by $t \geq 1$ customers with probability $(t - \alpha)/(n + \theta)$ or
- at an unoccupied table with probability $(m\alpha + \theta)/(n + \theta)$.

From Ewens's original derivation, θ is related to the mutation rate in the Wright–Fisher model and is therefore tied to the prior probability of observing a new species, or sitting at an unoccupied table, at the next stage. On the other hand, α reinforces the probability of observing new species in the future, given that we have observed a certain number of species so far. Thus, α affects the probability of observing new species and the probability of observing a specific species in opposite ways: when $\alpha < 0$ the number of blocks is bounded, so observing a new species decreases the number of unseen species and the future probability of observing new species but increases the probability of seeing the newly observed species again in the future; when $\alpha > 0$ the opposite is true; and when $\alpha = 0$ we are in the neutral sampling scheme considered by Ewens. Gnedin [44] has recently studied a different two-parameter model which allows for the possibility of a finite, but random, number of blocks.

5.3 Asymptotic properties

The extended parameter range of the two parameter model leads to different asymptotic regimes for various critical statistics of Ewens–Pitman partitions. For (α, θ) in the parameter space of (17), let $(\Pi_n)_{n\geq 1}$ be a collection of random partitions generated by the above two-parameter Chinese restaurant process. For each $n \geq 1$, let K_n denote the number of blocks of Π_n. When $\alpha < 0$ or $\alpha = 0$, the asymptotic behavior of K_n is clear from prior discussion: the $\alpha < 0$ case corresponds to Dirichlet-Multinomial sampling (Section 4.2), so that $K_n \to -\theta/\alpha$ a.s. as $n \to \infty$, while the $\alpha = 0$ case corresponds to Ewens sampling formula, whose description as a logarithmic combinatorial structure (Section 3.8) immediately gives $K_n \sim \theta \log(n)$ a.s. as $n \to \infty$. When $\alpha > 0$, Pitman [76] obtains the following limit law.

THEOREM 5.1 (Pitman [76]). *For $0 < \alpha < 1$ and $\theta > -\alpha$, $n^{-\alpha}K_n \to S_\alpha$ a.s., where S_α is a strictly positive random variable with continuous density*

$$\mathbb{P}\{S_\alpha \in dx\} = \frac{\Gamma(\theta + 1)}{\Gamma(\theta/\alpha + 1)} x^{\theta/\alpha} g_\alpha(x) dx, \quad x > 0,$$

and

$$g_\alpha(x) = \frac{1}{\pi\alpha} \sum_{k=0}^{\infty} \frac{(-1)^{k+1}}{k!} \Gamma(\alpha k + 1) \sin(\pi\alpha k) x^{k-1}, \quad x > 0,$$

is the Mittag–Leffler density.

The random variable S_α in Theorem 5.1 is called the *α-diversity* of $(\Pi_n)_{n\geq 1}$. Pitman goes on to characterize random partitions with a certain α-diversity in terms of the power law behavior of their relative block sizes [76, Lemma 3.11]. Of the many fascinating properties of the Ewens–Pitman family, the next description in terms of the jumps of subordinators provides some of the deepest connections to classical stochastic process theory.

5.4 Gamma and stable subordinators

For $0 \leq \alpha \leq 1$ and $\theta > 0$, the theory of Poisson–Kingman partitions [75] brings forth a remarkable connection between Ewens's sampling formula, the Poisson–Dirichlet distribution, and the jumps of subordinators. A *subordinator* $(\tau(s))_{s\geq 0}$ is an increasing stochastic process with stationary, independent increments whose distribution is determined by a *Lévy measure* Λ such that for each $s \geq 0$

$$\mathbb{E}(\exp\{-\lambda\tau(s)\}) = \exp\left(-s\int_0^\infty (1-\exp(-\lambda x))\Lambda(dx)\right).$$

From a random variable $T > 0$ and the closure Z of the range of $(\tau(s))_{s\geq 0}$, we define

$$V_1(T) \geq V_2(T) \geq \cdots \geq 0$$

as the ranked lengths of the subintervals within $[0,T] \setminus Z$. Thus, the ranked, normalized vector

$$\left(\frac{V_1(T)}{T}, \frac{V_2(T)}{T}, \ldots\right) \tag{18}$$

is a random element of $\mathcal{S}^{\downarrow}$ and its distribution must determine the law of an infinite exchangeable partition by Kingman's correspondence (Theorem 4.1).

Pitman & Yor [77] found several deep and explicit connections between the Poisson–Dirichlet distribution and the ranked vector in (18). For $\theta > 0$ and $b > 0$, let $(\tau(s))_{s\geq 0}$ be a *gamma subordinator* with Lévy measure

$$\Lambda(dx) = \theta x^{-1}e^{-bx}dx, \quad x > 0.$$

Then for all $\theta > 0$, the vector (18) with $T = \tau(\theta)$ has the Poisson–Dirichlet$(0,\theta)$ distribution. On the other hand, if $(\tau(s))_{s\geq 0}$ is an *α-stable subordinator* for $0 < \alpha < 1$, i.e., for some $C > 0$ the Lévy measure satisfies

$$\mathbb{E}(\exp\{-\lambda\tau(s)\}) = \exp\{-sC\Gamma(1-\alpha)\lambda^{\alpha}\},$$

then, for all $s > 0$, (18) with $T = \tau(s)$ has the Poisson–Dirichlet$(\alpha, 0)$ distribution. Other interesting special cases include the Poisson–Dirichlet(α,α) distribution, which arises as the ranked excursion lengths of semistable Markov bridges derived from α-stable surbordinators. In particular, the excursion lengths of Brownian motion on $[0,1]$ give rise to Poisson–Dirichlet$(1/2,0)$ and the excursion lengths of Brownian bridge on $[0,1]$ lead to Poisson–Dirichlet$(1/2,1/2)$. The Poisson–Dirichlet$(\alpha,0)$ distribution also arises in the low temperature asymptotics of Derrida's random energy model; see [25, 26] for further details.

In complete generality, Pitman & Yor [77, Proposition 21] derive the family of Poisson–Dirichlet measures for $0 < \alpha < 1$ and $\theta > 0$ by taking $(\tau(s))_{s\geq 0}$ to be a subordinator with Lévy measure $\Lambda(dx) = \alpha Cx^{-\alpha-1}e^{-x}dx$ and $(\gamma(t))_{t\geq 0}$ to be a gamma subordinator independent of $(\tau(s))_{s\geq 0}$. For $\theta > 0$, $S_{\alpha,\theta} = C^{-1}\gamma(\theta/\alpha)/\Gamma(1-\alpha)$, and $T = \tau(S_{\alpha,\theta})$, (18) has the Poisson–Dirichlet distribution with parameter (α,θ). Ishwaran & James [52, 53] make extensive use of Proposition 21 in their development of the generalized weighted Chinese restaurant process and Gibbs sampling methods for stick-breaking priors.

5.5 Size-biased sampling and the random allocation model

When studying mass partitions $(S_1, S_2, \ldots) \in \mathcal{S}^{\downarrow}$, it is sometimes more convenient to work with a size-biased reordering, denoted $\tilde{S} = (\tilde{S}_1, \tilde{S}_2, \ldots)$, in the infinite simplex

$$\mathcal{S} = \left\{ (s_1, s_2, \ldots) : \, s_i \geq 0, \, \sum_{k \geq 1} s_k \leq 1 \right\}.$$

Assuming $S \in \mathcal{S}^{\downarrow}$ satisfies $\sum_{k\geq 1} S_k = 1$, we obtain $\tilde{S}$ from $S = (S_1, S_2, \ldots)$ by putting $\tilde{S}_j = S_{I_j}$, where $I_1, I_2, \ldots$ are drawn randomly with distribution

$$\begin{aligned} \mathbb{P}\{I_1 = i \mid S\} &= S_i \quad \text{and} \\ \mathbb{P}\{I_j = i \mid S, \tilde{S}_1, \ldots, \tilde{S}_{j-1}\} &= \frac{S_i}{1 - \tilde{S}_1 - \cdots - \tilde{S}_{j-1}} \mathbf{1}\{\tilde{S}_1 \neq i, \ldots, \tilde{S}_{j-1} \neq i\}, \end{aligned}$$

as long as $\sum_{k=1}^{j-1} \tilde{S}_k < 1$. If $\sum_{k=1}^{j-1} \tilde{S}_k = 1$, we put $\tilde{S}_k = 0$ for all $k \geq j$. In words, we sample $I_1, I_2, \ldots$ without replacement from an urn with balls labeled $1, 2, \ldots$ and weighted $S_1, S_2, \ldots$, respectively. The distribution of a size-biased reordering of $S \sim$ Poisson–Dirichlet(α, θ) is called the *Griffiths–Engen–McCloskey distribution* with parameter (α, θ).

By size-biasing the frequencies of a Poisson–Dirichlet sequence, we arrive at yet another nifty construction in terms of the *random allocation model*, or *stick breaking process*. For example, let $U_1, U_2, \ldots$ be independent, identically distributed Uniform random variables on $[0, 1]$ and define $V_1, V_2, \ldots$ in $\mathcal{S}$ by

$$\begin{aligned} V_1 &= U_1, \\ V_2 &= U_2(1 - U_1), \quad \text{and} \\ V_k &= U_k(1 - U_{k-1}) \cdots (1 - U_1). \end{aligned}$$

Then $V = (V_1, V_2, \ldots)$ has the Griffiths–Engen–McCloskey distribution with parameter $(0, 1)$, i.e., V is distributed as a size-biased reordering of the block frequencies of a Ewens set partition with parameter $\theta = 1$. We can visualize the above procedure as a recursive breaking of a stick with unit length: we first break the stick U_1 units from the bottom; we then break the remaining piece $U_2(1 - U_1)$ units from the bottom; and so on.

Theorem 5.2 (Pitman [74]). *Let $V = (V_1, V_2, \ldots)$ be a size-biased reordering of components from the Poisson–Dirichlet distribution with parameter (α, θ). Then $V =_{\mathcal{D}} V^* = (V_1^*, V_2^*, \ldots)$, where*

$$\begin{aligned} V_1^* &= W_1, \\ V_2^* &= W_2(1 - W_1), \quad \textit{and} \\ V_k^* &= W_k(1 - W_{k-1}) \cdots (1 - W_1), \end{aligned}$$

for $W_1, W_2, \ldots$ independent and $W_j \sim$ Beta$(1 - \alpha, \theta + j\alpha)$ for each $j = 1, 2, \ldots$.

Note that $W_1, W_2, \ldots$ are identically distributed only if $\alpha = 0$, and so the stick breaking description further explains the self-similarity property of Ewens's distribution (Section 3.5). In particular, let $V = (V_1, V_2, \ldots)$ be distributed as a size-biased sample from a Poisson–Dirichlet mass partition with parameter $(0, \theta)$. With $V \setminus \{V_1\} = (V_2, V_3, \ldots)$, Theorem 5.2 implies that

$$(V_1, \frac{1}{1 - V_1} V \setminus \{V_1\}) =_{\mathcal{D}} (W, V), \tag{19}$$

where $W, V_1, V_2, \ldots$ are independent Beta$(1, \theta)$ random variables.

6. BAYESIAN NONPARAMETRICS AND CLUSTERING METHODS

6.1 Dirichlet process and stick breaking priors

Because of its many nice properties and convenient descriptions in terms of stick breaking, subordinators, etc., the Ewens–Pitman distribution (17) is widely applicable throughout statistical practice, particularly in the fast-developing field of Bayesian nonparametrics. In nonparametric problems, the parameter space is the collection of all probability distributions. As a practical matter, Bayesians often neglect subjective prior beliefs in exchange for a prior distribution whose "posterior distributions[...are] manageable analytically" [40, p. 209]. Conjugacy between the prior and posterior distributions in the Dirichlet-Multinomial process (Section 4.2) hints at a similar relationship in Ferguson's Dirichlet process prior [40] for nonparametric Bayesian problems. The construction of the Dirichlet process and Poisson–Dirichlet$(0, \theta)$ distributions from a gamma subordinator (Section 5.4) nails down the connection to Ewens's sampling formula.

A *Dirichlet process* with finite, non-null concentration measure β on $\mathcal{X}$ is a stochastic process S for which the random vector $(S(A_1), \ldots, S(A_k))$ has the $(k-1)$-dimensional Dirichlet distribution with parameter $(\beta(A_1), \ldots, \beta(A_k))$ for every measurable partition $A_1, \ldots, A_k$ of $\mathcal{X}$. Thus, a Dirichlet process S determines a random probability measure on $\mathcal{X}$ for which the conditional distribution of S, given $X_1, \ldots, X_n$, is again a Dirichlet process with concentration measure $\beta + \sum_{i=1}^{n} \delta_{X_i}$, where δ_{X_i} is a point mass at X_i. In particular, for a measurable partition $A_1, \ldots, A_k$, let N_j denote the number of points among $X_1, \ldots, X_n$ that fall in A_j for each $j = 1, \ldots, k$. Then the conditional distribution of $(S(A_1), \ldots, S(A_k))$ given $(N_1, \ldots, N_k)$ is Dirichlet with parameter $(\beta(A_1) + N_1, \ldots, \beta(A_k) + N_k)$, just as in Section 4.2.

From a realization $X_1, \ldots, X_n$ of the Dirichlet process, we can construct a partition Π_n as in (13). The posterior concentration measure $\beta + \sum_{i=1}^{n} \delta_{X_i}$ and basic properties of the Dirichlet distribution yield the update rule

$$\mathbb{P}\{X_{n+1} \in \cdot \mid X_1, \ldots, X_n\} = \frac{\beta(\cdot)}{n + \beta(\mathcal{X})} + \sum_{i=1}^{n} \frac{\delta_{X_i}(\cdot)}{n + \beta(\mathcal{X})},$$

from which the relationship to the Chinese restaurant rule with $\theta = \beta(\mathcal{X}) < \infty$ and Blackwell & MacQueen's [11] urn scheme is apparent.

More recently, the Ewens–Pitman distribution has been applied to sampling applications in fish trawling [78], analysis of rare variants [15], species richness in multiple populations [6], and Bayesian clustering methods [22]. In fact, ever since Ishwaran & James [52, 53] brought stick breaking priors and the Chinese restaurant process to the forefront of Bayesian methodology, the field of Bayesian nonparametrics has become one of the most active areas of statistical research.

Others [38, 65] have further contributed to the foundations of Bayesian nonparametrics laid down by Ishawaran & James. This overwhelming activity forbids any possibility of a satisfactory survey of the topic and promises to quickly outdate the contents of the present section. For a more thorough accounting of this rich area, we recommend other related work by the cited authors.

6.2 Clustering and classification

In classical and modern problems alike, statistical units often segregate into non-overlapping classes $B = \{B_1, B_2, \ldots\}$. These classes may represent the grouping of animals according to species, as in Fisher, Corbet & Williams's [42] and Good & Toulmin's [45] consideration of the number of unseen species in a finite sample, or the grouping of literary works according to author, as in Efron & Thisted's [30, 31] textual analysis of an unattributed poem from the Shakespearean era. While these past analyses employ parametric empirical Bayes [42] and nonparametric [45] models, modern approaches to machine learning and classification problems often involve partition models and Dirichlet process priors, e.g., [12, 69]. In many cases, the Ewens–Pitman family is a natural prior distribution for the true clustering $B = \{B_1, B_2, \ldots\}$.

For clustering based on categorical data sequences, e.g., DNA sequences, roll call data, and item response data, Crane [21] enlarges the parameter space of the Ewens–Pitman family to include an underlying clustering B. Given $\alpha_1, \ldots, \alpha_k > 0$, each block in B partitions independently according to the Dirichlet-Multinomial process with parameter $(\alpha_1, \ldots, \alpha_k)$. By ignoring the class labels as in (13), we obtain a new partition by aggregating across the blocks of B. The above procedure by first splitting within blocks and then aggregating across blocks warrants the description as a *cut-and-paste process* [18, 20]. When $\alpha_1 = \cdots = \alpha_k = \alpha > 0$, the cut-and-paste distribution amounts to

$$\mathbb{P}\{\Pi_n = \pi\} = k^{\downarrow \#\pi} \prod_{b \in B} \frac{\prod_{b' \in \pi} \alpha^{\uparrow \#(b \cap b')}}{(k\alpha)^{\uparrow \#b}}, \tag{20}$$

with the convention that $\alpha^{\uparrow 0} = 1$.

When $B = \{\{1, \ldots, n\}\}$, (20) coincides with the Dirichlet-Multinomial distribution in (16), equivalently Ewens–Pitman$(-\alpha, k\alpha)$ distribution in (17); and when $B = \{\{1\}, \ldots, \{n\}\}$, every individual chooses its block independently with probability $1/k$. In both cases, Π_n is exchangeable, but otherwise the distribution in (20) is only invariant under relabeling by permutations that fix B, a statistical property called *relative exchangeability*. In general, B is a partition of the population $\mathbb{N}$, but the marginal distribution of Π_n depends on B only through its restriction to $[n]$. Thus, the family is also sampling consistent and it enjoys the non-interference property (Section 3.6).

In the three-parameter model (20), the Ewens–Pitman prior for B exhibits nice properties and produces reasonable inferences [22]. For continuous response data, McCullagh & Yang [68] employ the *Gauss–Ewens cluster process*, whereby B obeys the Ewens distribution with parameter $\theta > 0$ and, given B, $(Y_1, Y_2, \ldots)$ is a sequence of multivariate normal random vectors with mean and covariance depending on B. Nice properties of the Gaussian and Ewens distributions combine to permit tractable calculation of posterior predictive probabilities and other quantities of interest for classification and machine learning applications.

7. INDUCTIVE INFERENCE

7.1 Rules of succession and the sufficientness postulate

Unbeknownst to Ewens, Ferguson, or Antoniak, the circle of ideas surrounding Ewens's sampling formula and the Dirichlet process prior lies at the heart of fundamental questions in inductive inference. Two centuries before Ewens's discovery, Bayes, Laplace, and De Morgan pondered epistemological questions about how past information can be used to update beliefs about the future [91]. For example, what is the probability the sun will rise tomorrow given that it has risen each of the previous N days? Laplace's famed *rule of succession*, which attributes probability $(N+1)/(N+2)$ to this event, follows from Bayes's [8] paradigm for "[events] concerning the probability of which we absolutely know nothing antecedently to any trials made concerning it." Under these circumstances, Bayes argued that he has "no reason" to assume anything other than a uniform prior on the possible outcomes, i.e., the number of successes S_n in n trials satisfies $\mathbb{P}\{S_n = k\} = 1/(n+1)$ for each $k = 0, 1, \ldots, n$. A straightforward mathematical argument [91] reveals that Bayes's principle of indifference implies a uniform prior distribution on the success probability of each outcome.

Johnson [54] later expanded upon Bayes's analysis by allowing for an event with possibly $k \geq 2$ different outcomes. In this case, the result of n trials is summarized by a vector $(n_1, \ldots, n_k)$, with n_i counting the number of outcomes of type $i = 1, \ldots, k$. Under Johnson's *sufficientness postulate*, by which the conditional probability that the $(n+1)$st observation is type i depends only on n_i and n, either all outcomes are independent or the conditional probabilities have the form of the Ewens–Pitman family with $\alpha < 0$ and $\theta = -k\alpha$; see Section 5.1 above. Just as Bayes's postulate implies the uniform prior, Johnson's sufficientness postulate implies the $(k-1)$-dimensional symmetric Dirichlet prior (15). Thus, Johnson, whose work predates Ferguson and Antoniak by forty years, provides a logical justification for the Dirichlet–Multinomial and Dirichlet process priors in Bayesian analysis.

At its core, Ewens's sampling formula is concerned with rules of succession, or more cavalierly, predicting the unpredictable [92]: Given an observed allelic partition $(m_1, m_2, \ldots)$, what is the probability that the next sampled individual is of a previously observed type or of a new type entirely? De Morgan [24] pondered this question more than a century before Ewens and arrived at a similar answer, but without any formal justification; see Section 4.4 above.

7.2 Zabell's universal continuum

A primary consideration of induction surrounds universal generalizations, e.g., what is the probability that the sun will rise tomorrow and every day in the future given that it has risen today and every day in the past? More generically, given that we have so far observed a partition $\Pi_n = \mathbf{1}_n = \{\{1, \ldots, n\}\}$ with all elements in the same block, what is the probability that we are actually sampling from the universal one-block partition $\mathbf{1}_\infty = \{\{1, 2, \ldots\}\}$? The Chinese restaurant process update probabilities implicitly entail the Poisson–Dirichlet process prior, which is absolutely continuous and assigns zero prior mass to the universal partition $\mathbf{1}_\infty$. Simply put, no amount of data is enough to nudge the posterior probability of $\mathbf{1}_\infty$ above zero.

Following the path of least resistance, Zabell [93] refines the Ewens–Pitman two-parameter family by taking ν in the paintbox process (14) to be the two-point mixture

$$\nu_{\alpha,\theta,\epsilon} = (1-\epsilon)\nu_{\alpha,\theta} + \epsilon\delta_{\mathbf{1}_\infty},$$

where $0 \le \epsilon \le 1$ is the prior probability assigned to the point mass $\delta_{\mathbf{1}_\infty}$ at the universal partition and $\nu_{\alpha,\theta}$ is the Poisson–Dirichlet measure with parameter (α,θ). Thus, $\epsilon = 0$ corresponds to the usual two-parameter family, and the conditional distribution of Π_{n+1}, given Π_n, coincides with the Chinese restaurant probabilities (Section 5.2) on the event $\Pi_n \neq \mathbf{1}_n$. On the event $\Pi_n = \mathbf{1}_n$, the above formulation assigns posterior probability

$$(1-\epsilon_n)\left(\frac{n-\alpha}{n+\theta}\right) + \epsilon_n$$

to the event that the $(n+1)$st observed species is the same type as all prior species and

$$(1-\epsilon_n)\left(\frac{\alpha+\theta}{n+\theta}\right)$$

to the event that the next species is new, where

$$\epsilon_n = \frac{\epsilon}{\epsilon + (1-\epsilon)\prod_{j=1}^{n-1}\frac{j-\alpha}{j+\theta}}$$

is the posterior probability of the event $\Pi_\infty = \mathbf{1}_\infty$ given $\Pi_n = \mathbf{1}_n$. Zabell's original derivation expresses these probabilities in terms of (α,θ,γ), with α and θ as before and $\gamma = (\alpha+\theta)\epsilon$.

8. COMBINATORICS, ALGEBRA, AND NUMBER THEORY

As if all the above instances were not enough, Ewens's sampling formula is also linked to two of the most fundamental ideas in mathematics, the determinant function in algebra and prime factorization in number theory.

8.1 The determinant and α-permanent

The *determinant* of an $n \times n$ matrix $M = (M_{i,j})_{1\le i,j\le n}$ is defined as

$$\det(M) = \sum_{\sigma:[n]\to[n]} \operatorname{sign}(\sigma) M_{1,\sigma(1)}M_{2,\sigma(2)}\cdots M_{n,\sigma(n)}, \tag{21}$$

where the sum is over all $n!$ permutations of $\{1,\dots,n\}$ and $\operatorname{sign}(\sigma)$ is the *parity* of σ, which equals $+1$ if σ is a product of an even number of cycles and equals -1 otherwise.

In the early 1800s, Cauchy [14] initiated the study of "fonctions symétriques permanentes," i.e., *permanent symmetric functions*, which ignore the parity of σ in (21). Cauchy's *permanent*

$$\operatorname{per}(M) = \sum_{\sigma:[n]\to[n]} M_{1,\sigma(1)}M_{2,\sigma(2)}\cdots M_{n,\sigma(n)} \tag{22}$$

resembles the determinant in appearance but little else: the determinant is easy to compute (e.g., as a product of eigenvalues), but the permanent is #P-complete

[83]; the determinant has a geometric interpretation in terms of volume, whereas the permanent's best interpretation is graph-theoretic.

Though determinant and permanent appear to occupy different mathematical territory, they come together in Vere-Jones's α-permanent [84]. For a complex-valued parameter α, the *α-permanent* of M is

$$\text{per}_\alpha(M) = \sum_{\sigma:[n]\to[n]} \alpha^{\text{cyc}(\sigma)} M_{1,\sigma(1)} M_{2,\sigma(2)} \cdots M_{n,\sigma(n)}, \tag{23}$$

where $\text{cyc}(\sigma)$ is the number of cycles of σ. Heuristically, (23) interpolates between (21) and (22), Cauchy's permanent (22) is the α-permanent with $\alpha = 1$ and the determinant (21) equals $(-1)^n \text{per}_{-1}(M)$, but its role is most prominent in modeling bosons and fermions in statistical physics [50, 66]. Amazingly, the α-permanent also incorporates Ewens's distribution in two different ways. (Note that the parameter α in (23) does not correspond directly to the parameter α in the Ewens–Pitman distribution (17).)

8.1.1 Random permutations As long as $\alpha > 0$ and $M_{i,j} > 0$ for all $i, j = 1, \ldots, n$, the α-permanent is the normalizing constant for the *cyclic product distribution* on permutations of $[n]$:

$$\mathbb{P}\{\Sigma_n = \sigma\} = \alpha^{\text{cyc}(\sigma)} \frac{M_{1,\sigma(1)} \cdots M_{n,\sigma(n)}}{\text{per}_\alpha(M)}. \tag{24}$$

Computational complexity of the α-permanent makes (24) intractable in general; however, when $M_{i,j} = 1$ for all i, j, we recover an exponential family of distributions on permutations,

$$\mathbb{P}\{\Sigma_n = \sigma\} = \frac{\alpha^{\text{cyc}(\sigma)}}{\alpha(\alpha+1)\cdots(\alpha+n-1)} = \exp\left\{\text{cyc}(\sigma)\log(\alpha) - \sum_{j=0}^{n-1}\log(\alpha+j)\right\}, \tag{25}$$

with natural parameter $\log(\alpha)$ and canonical sufficient statistic $\text{cyc}(\sigma)$. In the context of Section 4.5, (25) results by refining the Dubins–Pitman Chinese restaurant construction: the $(n+1)$st customer

- sits to the left of customer $j = 1, \ldots, n$ with probability $1/(\alpha+n)$ and
- sits alone at a table with probability $\alpha/(\alpha+n)$.

Occupied tables correspond to cycles in a random permutation and the left-to-right ordering of customers at each table determines the order of elements within each cycle. Just as in the Chinese restaurant construction in Section 4.5, we recover Ewens's distribution (6) from (25) by ignoring the order in which individuals are seated at each table. In particular, Σ_n induces a random partition Π_n whose distribution is the sum of (25) over all permutations with unordered cycles corresponding to the blocks of a specific partition of $[n]$.

Developed in this way, Ewens's distribution is a subfamily of the *cyclic product distribution* on partitions of $[n]$. For any subset $b \subseteq [n]$, the sum of cyclic products

$$\text{cyp}(M)[b] = \sum_{\sigma:b\to b \text{ s.t. } \text{cyc}(\sigma)=1} \prod_{i\in b} M_{i,\sigma(i)}$$

is the sum over all permutations of b with a single cycle, and so the α-permanent decomposes into a sum over partitions of $[n]$ by

$$\mathrm{per}_\alpha(M) = \sum_\pi \alpha^{\#\pi} \prod_{b\in\pi} \mathrm{cyp}(M)[b].$$

The (α, M)-*cyclic product distribution* has the form of a product partition model,

$$\mathbb{P}\{\Pi_n = \pi\} = \alpha^{\#\pi} \frac{\prod_{b\in\pi} \mathrm{cyp}(M)[b]}{\mathrm{per}_\alpha(M)} \propto \prod_{b\in\pi} \alpha \cdot \mathrm{cyp}(M)[b];$$

see Section 3.7.

8.1.2 The two-parameter model The permanental decomposition theorem [19] expresses the α-permanent as a sum over partitions of $[n]$ of permanents of related matrices. In particular, for real constants α and β,

$$\mathrm{per}_{\alpha\beta}(M) = \sum_\pi \beta^{\downarrow\#\pi} \prod_{b\in\pi} \mathrm{per}_\alpha(M[b]), \tag{26}$$

where $M[b] = (M_{i,j})_{i,j\in b}$ is the submatrix of M whose rows and columns are indexed by $b \subseteq [n]$ and the sum is over all partitions of $[n]$. As long as every term in (26) is nonnegative, we obtain the *permanental partition model*

$$\mathbb{P}\{\Pi_n = \pi\} = \beta^{\downarrow\#\pi} \frac{\prod_{b\in\pi} \mathrm{per}_\alpha(M[b])}{\mathrm{per}_{\alpha\beta}(M)}, \tag{27}$$

where $M_{i,j}$ is a measure of similarity between elements i and j and $\beta^{\downarrow n} = \beta(\beta - 1)\cdots(\beta - n + 1)$. In a homogeneous environment, i.e., $M_{i,j} \equiv 1$, (27) becomes

$$\mathbb{P}\{\Pi_n = \pi\} = \beta^{\downarrow\#\pi} \frac{\prod_{b\in\pi} \alpha^{\uparrow\#b}}{(\alpha\beta)^{\uparrow n}},$$

which equals the Ewens–Pitman distribution (17) under the substitution $\alpha \mapsto -\alpha$ and $\beta \mapsto \theta/\alpha$. In this way, the permanental partition model (27) extends the Ewens–Pitman two-parameter family to a three-parameter distribution, but (27) is neither exchangeable nor consistent in general.

8.2 Random numbers and large prime factors

For $n = 1, 2, \ldots$, let N_n be uniformly distributed in $\{1, \ldots, n\}$, i.e.,

$$\mathbb{P}\{N_n = i\} = 1/n, \quad i = 1, \ldots, n.$$

By the fundamental theorem of arithmetic, N_n can be factored uniquely into a product of prime numbers, i.e., there exists a unique sequence $P_{n,1} \geq \cdots \geq P_{n,k} \geq 2$ of primes such that

$$N_n = P_{n,1} \times \cdots \times P_{n,k}.$$

Since N_n is random, so is $P_n^\downarrow = (P_{n,1}, \ldots, P_{n,k})$. Moreover, we can express

$$\log(N_n) = \log(P_{n,1}) + \cdots + \log(P_{n,k})$$

so that the normalized vector

$$S_n^{\downarrow} = \left(\frac{\log(P_{n,1})}{\log(n)}, \dots, \frac{\log(P_{n,k})}{\log(n)}, 0, 0, \dots\right) \tag{28}$$

is a random element of the ranked-simplex $\mathcal{S}^{\downarrow}$.

Billingsley [10] first studied the distribution of the ranked, normalized prime factors $S_n^{\downarrow}$. Donnelly & Grimmett [29] followed twenty years later with a simpler proof based on the size-biased reordering $\tilde{S}_n$ of $S_n^{\downarrow}$. In light of all previous discussion, their conclusion is astonishing: the relative sizes of the prime factors of a uniform random integer converge in distribution to the asymptotic block sizes of a Ewens partition with $\theta = 1$.

Theorem 8.1 (Billingsley [10]; Donnelly & Grimmett [29]). *For each $n = 1, 2, \dots$, let $P_n^{\downarrow} = (P_{n,1}, \dots, P_{n,k})^{\downarrow}$ be the prime factorization of a uniform random integer in $\{1, \dots, n\}$, let $S_n^{\downarrow}$ be the normalized, ranked vector in (28), and let $\tilde{S}_n$ be its size-biased reordering. Then*

$$S_n^{\downarrow} \longrightarrow_{\mathcal{D}} \textit{Poisson–Dirichlet}(0,1) \quad \textit{or, equivalently,} \tag{29}$$

$$\tilde{S}_n \longrightarrow_{\mathcal{D}} \textit{Griffiths–Engen–McCloskey}(0,1). \tag{30}$$

Knuth & Trabb Pardo [64] and Vershik [85] show the same distributional convergence for N_n drawn uniformly in $\{n, n+1, \dots, 2n\}$ and N_n drawn from the *Riemann zeta distribution*,

$$\mathbb{P}\{N_n = x\} = x^{-s_n}/\zeta(s_n), \quad x = 1, 2, \dots,$$

where $\zeta(s) = \sum_{x=1}^{\infty} x^{-s_n}$ and $s_n = 1 + 1/\log(n)$.

8.3 Macdonald polynomials

Recent work at the interface of representation theory, algebraic combinatorics, and probability has led to interesting connections between symmetric polynomials and fundamental notions in combinatorial stochastic process theory, particularly Kingman's theorem [57, 71]. *Macdonald processes* [13] are a particularly interesting family of probability distributions on sequences of integer partitions that arise in certain models for interacting particle systems and directed random polymers in statistical physics. The Macdonald process is named after its description in terms of Macdonald polynomials [67], a family of orthogonal polynomials with two parameters (q, t). Macdonald polynomials generalize various other families of symmetric polynomials, e.g., Hall–Littlewood and Jack polynomials, and thus arise throughout representation theory and algebraic combinatorics. Within probability theory, Diaconis & Ram [27] have recently provided an interpretation in terms of the stationary distribution of a special Markov chain on spaces of integer partitions. Ewens's sampling formula (1) arises as the limit of this stationary distribution under the regime $q = t^{1/\theta}$ and $t \to 1$, for $\theta > 0$. See [13] and [27, Section 2.4.2] for further details.

9. CONCLUDING REMARKS

A confluence of mathematical, statistical, and scientific facts contributes to the ubiquity of Ewens's sampling formula: Ewens's initial assumptions of neutral mutation and independence between individuals suited a need for a tractable mathematical theory of allele sampling; Laplace's rule of succession, De Morgan's urn scheme, and Johnson's sufficientness postulate all arise from principles of indifference at the heart of Bayesian epistemology [23]; Ferguson [40] and Antoniak [2] stumbled upon Ewens's sampling formula without regard for the above logical properties or principles of inductive inference; and the same mathematical properties that drive Ferguson's and Antoniak's approach underlie the deep connections between Ewens's sampling formula and classical stochastic process theory via the Poisson–Dirichlet distribution [39, 76, 77]. These discoveries along with the occurrence of Ewens's sampling formula in the realm of matrix permanents [19] and prime divisors [10, 29] hint at deep roots in the foundations of mathematics. Still new uses of the sampling formula in clustering problems [21, 22, 68, 69] manifest its utility in modern applications. Altogether, Ewens's sampling formula envelops a rich history of important contributions within classical and modern scientific, mathematical, and statistical study.

As much as space permits, the foregoing survey provides a comprehensive modern overview of Ewens's sampling formula. Less prominent but equally intriguing connections to rumor spreading [7], physics [48], computation [70], and many other areas are scattered throughout the literature. If recent trends in Bayesian statistics and stochastic process theory are any indication, a book length monograph will soon be necessary to adequately summarize the varied occurrences of Ewens's sampling formula.

REFERENCES

[1] D. J. Aldous. Exchangeability and related topics. In *École d'été de probabilités de Saint-Flour, XIII—1983*, volume 1117 of *Lecture Notes in Math.*, pages 1–198. Springer, Berlin, 1985.

[2] C. Antoniak. Mixtures of dirichlet processes with applications to nonparametric problems. *The Annals of Statistics*, 2:1152–1174, 1974.

[3] R. Arratia, A. Barbour, and S. Tavaré. Poisson process approximations for the Ewens sampling formula. *Annals of Applied Probability*, 2(3):519–535, 1992.

[4] R. Arratia, A. Barbour, and S. Tavaré. Limits of Logarithmic Combinatorial Structures. *Annals of Probability*, 28(4):1620–1644, 2000.

[5] R. Arratia, A. Barbour, and S. Tavaré. *Logarithmic Combinatorial Structures: a Probabilistic Approach*. European Mathematical Society, 2003.

[6] S. Bacallado, S. Favaro, and L. Trippa. Bayesian nonparametric inference for shared species richness in multiple populations. *Journal of Statistical Planning and Inference*, 2014.

[7] D. Bartholomew. *Stochastic Models for Social Processes, second edition*. John Wiley & Sons, London, 1973.

[8] T. Bayes. An essay toward solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53:370–418, 1764.

[9] J. Bertoin. *Random fragmentation and coagulation processes*, volume 102 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2006.

[10] P. Billingsley. On the distribution of large prime factors. *Period. Math. Hungar.*, 2:283–289, 1972.

[11] D. Blackwell and J. MacQueen. Ferguson distributions via pólya urn schemes. *Annals of Statistics*, 1:353–355, 1973.

[12] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[13] A. Borodin and I. Corwin. Macdonald Processes. *Probability Theory and Related Fields*, 158:225–400, 2014.

[14] A. Cauchy. Mémoire sur les fonctions qui ne peuvent obtenir que deux valeurs égales et de signes contraires par suite des transpositions opérées entre les variables quelles renferment. *Journal de l'École Polytechnique*, 10:91–169, 1815.

[15] O. Cesari, S. Favaro, and B. Nipoti. Posterior analysis of rare variants in Gibbs-type species sampling models. *Journal of Multivariate Analysis*, 131:79–98, 2014.

[16] D. Champernowne. A model of income distribution. *Econ. J.*, 63(318), 1953.

[17] F. Christiansen. *Theories of Population Variation in Genes and Genomes*. Princeton University Press, 2014.

[18] H. Crane. A consistent Markov partition process generated from the paintbox process. *J. Appl. Probab.*, 43(3):778–791, 2011.

[19] H. Crane. Some algebraic identities for the α-permanent. *Linear Algebra and Its Applications*, 439(11):3445–3459, 2013.

[20] H. Crane. The cut-and-paste process. *Annals of Probability*, 42(5):1952–1979, 2014.

[21] H. Crane. Clustering from categorical data sequences. *Journal of the American Statistical Association*, 110(510):810–823, 2015.

[22] H. Crane. Generalized Ewens–Pitman model for Bayesian clustering. *Biometrika*, 102(1):231–238, 2015.

[23] B. de Finetti. La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, 7:1–68.

[24] A. De Morgan. *An Essay on Probabilities, and on their Application to Life Contingencies and Insurance Offices*. Longman, et al, London, 1838.

[25] B. Derrida. Random-energy model: an exactly solvable model of disordered systems. *Phys. Rev. B*, 24(5):2613–2626, 1981.

[26] B. Derrida. From random walks to spin glasses. *Phys. D*, 107(2-4):186–198, 1997.

[27] P. Diaconis and A. Ram. A probabilistic interpretation of the Macdonald polynomials. *The Annals of Applied Probability*, 40(5):1861–1896, 2012.

[28] P. Donnelly. Partition Structures, Polya Urns, the Ewens Sampling Formula, and the Ages of Alleles. *Theoretical Population Biology*, 30:271–288, 1986.

[29] P. Donnelly and G. Grimmett. On the asymptotic distribution of large prime factors. *J. London Math. Soc*, 47:395–404, 1993.

[30] B. Efron and R. Thisted. Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika*, 63:435–447, 1976.

[31] B. Efron and R. Thisted. Did Shakespeare write a newly discovered poem? *Biometrika*, 74:445–455, 1987.

[32] S. Ethier and R. Griffiths. The transition function of a Fleming–Viot process. *The Annals of Probability*, 21(3):1571–1590, 1993.

[33] R. Etienne. A new sampling formula for neutral biodiversity. *Ecology Letters*, 8:253–260, 2005.

[34] R. Etienne. A neutral sampling formula for multiple samples and an 'exact' text of neutrality. *Ecology Letters*, 10:608–618, 2007.

[35] R. Etienne and D. Alonso. A dispersal-limited sampling theory for species and alleles. *Ecology Letters*, 8(11):1147–1156, 2005.

[36] W. Ewens and S. Tavaré. The Ewens sampling formula. In *Encyclopedia of Statistical Science, Eds. Kotz, S., Read, C.B. and Banks, D.L.* Wiley, New York, 1998.

[37] W. J. Ewens. The sampling theory of selectively neutral alleles. *Theoret. Population Biology*, 3:87–112, 1972.

[38] S. Favaro, A. Lijoi, and I. Pruenster. Conditional formulae for Gibbs-type exchangeable random partitions. *The Annals of Applied Probability*, 23:1721–1754, 2013.

[39] S. Feng. *The Poisson-Dirichlet Distribution and Related Topics*. Probability and its Applications. Springer-Verlag, Berlin, 2010.

[40] T. Ferguson. A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1(2):209–230, 1973.

[41] R. Fisher. On the dominance ratio. *Proceedings of the Royal Society of Edinburgh*, 42:321–341, 1922.

[42] R. Fisher, A. Corbet, and C. Williams. The relation between the number of species and the number of individuals in a random sample of an animal population. *The Journal of Animal Ecology*, 12:42–58, 1943.

[43] W. Fleming and M. Viot. Some Measure-Valued Markov Processes in Population Genetics Theory. *Indiana University Mathematics Journal*, 28(5):817–843, 1979.

[44] A. Gnedin. A species sampling model with finitely many types. *Electronic Communications in Probability*, 15:79–88, 2010.

[45] I. Good and G. Toulmin. The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika*, 43:45–63, 1956.

[46] R. Griffiths. Exact sampling distributions from the infinite neutral alleles model. *Advances in Applied Probability*, 11:326–354, 1979.

[47] J. A. Hartigan. Partition models. *Comm. Statist. Theory Methods*, 19(8):2745–2756, 1990.

[48] P. Higgs. Frequency distributions in population genetics parallel those in statistical physics. *Physical Review E*, 51:1–7, 1995.

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

[50] J. B. Hough, M. Krishnapur, Y. Peres, and B. Virág. Determinantal Processes and Independence. *Probability Surveys*, 3:206–229, 2006.

[51] S. Hubbell. *The Unified Neutral Theory of Biodiversity and Biogeography*. Princeton University Press, 2001.

[52] H. Ishwaran and L. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96:161–173, 2001.

[53] H. Ishwaran and L. James. Generalized weighted Chinese restaurant processes for species sampling mixture models. *Statistica Sinica*, 13:1211–1235, 2003.

[54] W. Johnson. Probability: The deductive and inductive problems. *Mind*, 41:409–423, 1932.

[55] S. Karlin and J. McGregor. Addendum to a paper of W. Ewens. *Theoretical Population Biology*, 3:113–116, 1972.

[56] S. Kerov. Coherent random allocations, and the ewens-pitman formula. *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI)*, 325 (Teor. Predst. Din. Sist. Komb.i Algoritm. Metody. 12)(246):127–145, 2005.

[57] S. Kerov, A. Okounkov, and G. Olshanski. The boundary of the Young graph with Jack edge multiplicities. *International Mathematics Research Notices*, 4:173–199, 1998.

[58] M. Kimura. Evolutionary rate at the molecular level. *Nature*, 217:624–6, 1968.

[59] J. Kingman. The population structure associated with the Ewens sampling formula. *Theoretical Population Biology*, 11:274–283, 1977.

[60] J. F. C. Kingman. Random partitions in population genetics. *Proc. Roy. Soc. London Ser. A*, 361(1704):1–20, 1978.

[61] J. F. C. Kingman. The representation of partition structures. *J. London Math. Soc. (2)*, 18(2):374–380, 1978.

[62] J. F. C. Kingman. *Mathematics of genetic diversity*, volume 34 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, Pa., 1980.

[63] J. F. C. Kingman. The coalescent. *Stochastic Process. Appl.*, 13(3):235–248, 1982.

[64] D. Knuth and L. Trabb Pardo. Analysis of a simple factorization algorithm. *Theoret. Comput. Sci.*, 3(3):321–348, 1976/77.

[65] A. Lijoi, R. Mena, and I. Pruenster. Bayesian nonparametric estimation of the probability of discovering a new species. *Biometrika*, 94:769–786, 2007.

[66] O. Macchi. The coincidence approach to stochastic point processes. *Advances in Applied Probability*, 7(1):83–122, 1975.

[67] I. Macdonald. *Symmetric Functions and Hall Polynomials, Second Edition*. Oxford Mathematical Monographs. Clarendon Press, Oxford, 1995.

[68] P. McCullagh and J. Yang. Stochastic classification models. In *International Congress of Mathematicians. Vol. III*, pages 669–686. Eur. Math. Soc., Zürich, 2006.

[69] P. McCullagh and J. Yang. How many clusters? *Bayesian Anal.*, 3(1):101–120, 2008.

[70] R. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9:249–265, 2000.

[71] G. Olshanski. Random permutations and related topics. *arXiv:1104.1266v2*, 2011.

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

[73] J. Pitman. Exchangeable and partially exchangeable random partitions. *Probab. Theory Related Fields*, 102(2):145–158, 1995.

[74] J. Pitman. Random discrete distributions invariant under size-biased permutation. *Adv. in Appl. Probab.*, 28(2):525–539, 1996.

[75] J. Pitman. Poisson-Kingman partitions. In *Statistics and science: a Festschrift for Terry Speed*, volume 40 of *IMS Lecture Notes Monogr. Ser.*, pages 1–34. Inst. Math. Statist., Beachwood, OH, 2003.

[76] J. Pitman. *Combinatorial stochastic processes*, volume 1875 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2006.

[77] J. Pitman and M. Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann. Probab.*, 25(2):855–900, 1997.

[78] M. Sibuya. Prediction in Ewens–Pitman sampling formula and random samples from number partitions. *Annals of the Institute of Statistical Mathematics*, 66:833–864, 2014.

[79] H. Simon. On a class of skew distribution functions. *Biometrika*, 42:425–440, 1955.

[80] N. Sloane. Online Encyclopedia of Integer Sequences. *Published electronically at http://www.oeis.org/*.

[81] R. Spielman, R. McGinnis, and W. Ewens. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *American Journal of Human Genetics*, 52(3):506–516, 1993.

[82] S. Tavaré and W. Ewens. The Multivariate Ewens Distribution. In *Discrete Multivariate Distributions (N.L. Johnson, S. Kotz and N. Balakrishnan Eds.)*. Wiley, 1997.

[83] L. G. Valiant. The complexity of computing the permanent. *Theoretical Computer Science*, 8:189–201, 1979.

[84] D. Vere-Jones. A generalization of permanents and determinants. *Linear Alg. Appl.*, 111:119–124, 1988.

[85] A. Vershik. Asymptotic distribution of decompositions of natural numbers into prime divisors (Russian). *Dokl. Akad. Nauk SSSR*, 289(2):269–272, 1986.

[86] J. Wakeley. *Coalescent Theory: An Introduction*. Roberts and Company Publishers, 2008.

[87] G. Watterson. Heterosis or neutrality? *Genetics*, 85:789–814, 1977.

[88] G. Watterson. The homozygosity test of neutrality. *Genetics*, 88:405–417, 1978.

[89] S. Wright. Evolution in Mendelian populations. *Genetics*, 16:97–159.

[90] G. Yule. A mathematical theory of evolution, based on the conclusions of dr. j.c. willis, f.r.s. *Phil. Trans. Roy. Soc. London, B*, 213:21–87, 1925.

[91] S. Zabell. Symmetry and Its Discontents. In *Causation, Chance, and Credence*, volume 1, pages 155–190. Kluwer Academic Publishers, 1988.

[92] S. Zabell. Predicting the Unpredictable. *Synthese*, 90(2):205–232, 1992.

[93] S. Zabell. The Continuum of Inductive Methods Revisited. In *The Cosmos of Science: Essays of Exploration*. The University of Pittsburg Press, 1997.