

RANDOM PERMUTATIONS AND THE EWENS SAMPLING FORMULA IN GENETICS

G. J. BABU¹

Department of Statistics, 319 Thomas Building, The Pennsylvania State University,
University Park, PA 16802, USA

E. MANSTAVIČIUS²

Department of Mathematics, Vilnius University, Naugarduko str. 24,
LT 2006 Vilnius, Lithuania

ABSTRACT

In the last few decades, mathematical population geneticists have been exploring the mechanisms that maintain diversity in a population. In 1972, Ewens established a formula to describe the probability distribution of a sample of genes from a population that has evolved over many generations, by a family of measures on the set of partitions of an integer. The Ewens formula can be used to test if the popular assumptions are consistent with data, and to estimate the parameters. The statistics that are useful in this connection will generally be expressed as functions of the sums of transforms of the *allelic partition*. Such statistics can be viewed as functions of a process on the permutation group of integers. The Ewens sampling formula also arises in Bayesian statistics via Dirichlet processes. Necessary and sufficient conditions for a process defined through the Ewens sampling formula to converge in a functional space to a stable process are presented. A counter example to show that these conditions are not necessary for one-dimensional convergence is constructed.

1. INTRODUCTION

Mathematical *population geneticists* have been exploring, for the past few decades, the mechanisms that maintain genetic diversity in a population. Natural *selection* due to interaction with the environment and *mutation* are some of the factors that influence genetic evolution. Some geneticists believe that the mutation and random fluctuations that are inherent in the reproductive process account for much of the genetic diversity. They view that the effect of selection has been exaggerated and hence concentrate on the so called *neutral alleles* models. It is the property of this model that there is no meaningful way of labelling the alleles.

¹Research supported in part by NSA grant MDA904-97-1-0023, NSF grant DMS-9626189, and by National Research Council's 1997-99 Twinning Fellowship.

²Research supported in part by Lithuanian Science and Studies Fund and by National Research Council's 1997-99 Twinning Fellowship.

The study of sampling distribution of genes from a population, that has evolved over many generations, helps in understanding the genetic structure of the population and in estimating the mutation rates.

2. THE EWENS SAMPLING FORMULA

Ewens (1972) established a formula to describe the sampling distribution of a sample of n genes from a large population by a partly heuristic argument. In several genetic models it is an exact formula and in others it is a close approximation. The formula is derived under the null hypothesis that there is no selection. In this case the *allelic partition*, $\bar{k} = (k_1, \dots, k_n)$, contains all the information available in a sample of n genes, where k_j denotes the number of alleles represented j times in the sample, $j = 1, \dots, n$. In other words Ewens formula provides the distribution of the multiplicities of alleles of a sample of genes from the so-called neutral alleles model of population genetics.

The Ewens sampling formula (Ewens, 1972) is given by

$$v_{n,\theta}\{(k_1, \dots, k_n)\} = \frac{n!}{\theta_{(n)}} \prod_{j=1}^n \left(\frac{\theta}{j}\right)^{k_j} \frac{1}{k_j!}, \quad (1)$$

where $\theta > 0$, $\theta_{(n)} = \theta(\theta + 1) \dots (\theta + n - 1)$, $k_j \geq 0$, and

$$1k_1 + \dots + nk_n = n. \quad (2)$$

The vector \bar{k} represents a partition of the integer n . The derivation of equation (1) was made rigorous by Karlin and McGregor (1972). Kingman (1980) argues that several different approaches lead to (1), under very broad assumptions. He claims that "The formula is reliable when

- (a) the size of the population is large compared to n , and the expected total number of mutations per generation is moderate (differing from θ only by a numerical factor),
- (b) the population is in statistical equilibrium under mutation and genetic drift, with selection at the locus playing a negligible role, and
- (c) mutation is nonrecurrent, so that every mutant allele is a completely novel one."

The Ewens sampling formula exposed the inadequacy of the 'standard' methods for estimation of mutation rates. It shows that \bar{k} is a sufficient statistic for θ , so that estimation of θ by using the sample heterozygosity (which is the least informative part of the data) rather than by \bar{k} is inefficient. The formula can be used to test if the assumptions (a), (b), and (c) are consistent with the data, and to estimate the parameter θ . The statistics that are of interest can generally be expressed as functions of the sums $\sum_{j \leq r} h_j(k_j)$, where $r \geq 1$ and h_j is a function on the set of non-negative integers. While the sampling theory of neutral alleles is still developing, the focus has shifted more toward DNA sequence data in recent years.

2.1. Ewens formula in Bayesian nonparametric problems

The Ewens formula also made its impact in an entirely different field, Bayesian statistics. It is well known that *Dirichlet processes* play an important role in Bayesian approach to nonparametric problems. A random probability measure D on a measure space (Ω, \mathcal{A}) with parameter α is called a Dirichlet process, if for every $k = 1, 2, \dots$ and measurable partition A_1, \dots, A_k of Ω , the joint distribution of $(D(A_1), \dots, D(A_k))$ is Dirichlet with parameters $(\alpha(A_1), \dots, \alpha(A_k))$. Antoniak (1974) has shown that Ewens formula can be used to test if a given data is from a Dirichlet process with unknown parameter α , using the actual pattern of multiplicities observed. Suppose a sample of size n from a Dirichlet process with parameter α is drawn. If α is *nonatomic*, then Antoniak (1974) establishes that the probability that the sample contains k_j elements that occur exactly j times, $j = 1, \dots, n$, is given by the Ewens formula (1), with $\theta = \alpha(\Omega)$.

3. GROUP OF PERMUTATIONS

Another interesting aspect of the Ewens formula is its combinatorial content. It is central to the study of a broad class of combinatorial structures such as permutation groups. Ewens formula can be viewed as a measure on the space of partitions of an integer n . A brief description of permutation groups and associated conjugate elements is presented here.

Let S_n denote the symmetric group of permutations on $\{1, \dots, n\}$. The elements of S_n can be represented uniquely by the product of independent cycles. More precisely, let $\sigma \in S_n$ be an arbitrary permutation and

$$\sigma = \kappa_1 \cdots \kappa_w, \quad (3)$$

be its unique representation (up to the order) by the product of independent cycles κ_i , where $w = w(\sigma)$ denotes the number of cycles. For example, when $n = 8$, the permutation τ that maps $\{1, 2, 3, 4, 5, 6, 7, 8\}$ onto $\{5, 3, 6, 1, 8, 2, 7, 4\}$ (i.e., $\tau(1) = 5, \tau(2) = 3, \tau(3) = 6, \tau(4) = 1, \tau(5) = 8, \tau(6) = 2, \tau(7) = 7, \tau(8) = 4$), has three cycles $(1\ 5\ 8\ 4)$, $(2\ 3\ 6)$ and (7) . The length of the first cycle is four, the length of the second cycle is three and the last one has length one. So one can write

$$\tau = (1\ 5\ 8\ 4)(2\ 3\ 6)(7).$$

Similarly, $\tau^2 = (1\ 8)(4\ 5)(2\ 3\ 6)(7)$, $\tau^3 = (1\ 4\ 8\ 5)(2)(3)(6)(7)$.

The order $\text{Ord}(\sigma)$ of a permutation σ is defined as the least positive integer k for which $\sigma^k = I$, where I is the identity permutation. For the example above, $\text{Ord}(\tau) = 12$. It is well known that $\text{Ord}(\sigma)$ is the least common multiple of $\{j : k_j(\sigma) > 0\}$. The asymptotic distribution of the order function was studied by Erdős and Turán (1965, 1967). They established that

$$\frac{1}{n!} \# \left(\sigma \in S_n : \log \text{Ord}(\sigma) - 0.5 \log^2 n \leq \frac{y}{\sqrt{3}} \log^{3/2} n \right) \rightarrow \Phi(y),$$

as $n \rightarrow \infty$, where Φ denotes the standard normal distribution function. The function $\log \text{Ord}$ is closely related to the function $\sum^* \log j$, where the sum \sum^* is extended over all positive $k_j(\sigma)$.

For $\sigma \in S_n$ the representation (3) leads to $1k_1(\sigma) + \dots + nk_n(\sigma) = n$, where $k_j(\sigma)$ denotes the number of cycles of σ of length j . The group S_n can be partitioned into equivalence classes by identifying the elements σ by the vector $\bar{k} = (k_1, \dots, k_n)$, where $0 \leq k_j = k_j(\sigma) \leq n$ and $1k_1 + \dots + nk_n = n$. In this case $w(\sigma) = k_1 + \dots + k_n$, the total number of cycles of σ , describes an additive function on S_n .

For each $\theta > 0$, the Ewens formula can be considered as a measure on the symmetric group S_n of permutations on $\{1, \dots, n\}$. This motivates a study of the distribution of values of a function on S_n . A functional limit theorem for a partial sum process $\sum_{j \leq y(t)} h_j(k_j)$, $0 \leq t \leq 1$, under the Ewens sampling formula, is described in this paper, where y is a function on the unit interval and h_j are functions on the set of non-negative integers satisfying $h_j(0) = 0$. The derivations involve concepts and ideas from probabilistic number theory.

3.1. Functions on permutations group

Let G be an additive abelian group. A map $h: S_n \rightarrow G$ is called an additive function if it satisfies the relation

$$h(\sigma) = \sum_{j=1}^n h_j(k_j(\sigma)) \quad (4)$$

for each $\sigma \in S_n$, where $h_j(0) = 0$ and $h_j(k)$, $j \geq 1$, $k \geq 0$, is some double sequence in G . If $h_j(k) = kh_j(1)$ for each $1 \leq j \leq n$ and $k \geq 0$, then h is called completely additive function. The number of cycles $w(\sigma)$ in (3) is a typical example of a completely additive function.

Similarly, a complex valued function f on S_n , given by $f(\sigma) = \prod_{j=1}^n f_j(k_j(\sigma))$ with $f_j(0) = 1$ is called multiplicative. It is called completely multiplicative if, in addition, $f_j(k) = f_j(1)^k$ holds for each $j \geq 1$ and $k \geq 0$. All these functions are measurable with respect to the finite field \mathcal{F} of subsets of S_n generated by the system of conjugate classes. For each $\theta > 0$, $\nu_{n,\theta}$ defined in (1) is a probability measure on \mathcal{F} . The uniform distribution (the Haar measure) on S_n induces the measure $\nu_{n,1}$ on the space of conjugate classes. This is the Ewens formula when $\theta = 1$.

4. FUNCTIONAL LIMIT THEOREM

The main result is based on the well known relation

$$\nu_{n,\theta}(\bar{k}) = P(\xi_1 = k_1, \dots, \xi_n = k_n \mid \zeta = n), \quad (5)$$

where ξ_j , $1 \leq j \leq n$ are independent Poisson random variables, satisfying $E\xi_j = \theta/j$, $\zeta = 1\xi_1 + \dots + n\xi_n$ (see, for instance, Arratia *et al.*, 1992).

To state the general invariance principle for additive functions $h(\sigma)$ given by (4), set for brevity $a(j) = h_j(1)$, and $u^* = (1 \wedge |u|)\text{sgn } u$, where $a \wedge b := \min\{a, b\}$. Here and in what follows we take the limits as $n \rightarrow \infty$. Let the normalizing factors $\beta(n) > 0$ satisfy $\beta(n) \rightarrow \infty$. The sequence $\{\beta(n)\}$ need not be monotone. Define

$$B(u, n; h) = \sum_{j \leq u} \left(\frac{a(j)}{\beta(n)} \right)^{*2} \frac{1}{j}, \quad A(u, n; h) = \theta \sum_{j \leq u} \left(\frac{a(j)}{\beta(n)} \right)^{*} \frac{1}{j}$$

and $y(t) := y_n(t) = \max\{l \leq n : B(l, n; h) \leq tB(n, n; h)\}$, $t \in [0, 1]$. We shall consider the weak convergence (denoted by \Rightarrow) of the process

$$H_{n,h} := H_{n,h}(\sigma, t) = \frac{1}{\beta(n)} \sum_{j \leq y(t)} h_j(k_j(\sigma)) - A(y(t), n; h), \quad t \in [0, 1]$$

under the measure $\nu_{n,\theta}$, in the space $\mathbf{D}[0, 1]$ endowed with the Skorohod topology (Billingsley, 1968). The corresponding process X_n with independent increments is defined by

$$X_{n,h} := X_{n,h}(t) = \frac{1}{\beta(n)} \sum_{j \leq y(t)} a(j)\xi_j - A(y(t), n; h), \quad t \in [0, 1]. \quad (6)$$

We consider the weak convergence (denoted by \Rightarrow) of the process $H_{n,h}$. Babu and Manstavičius (1998b) obtained the following theorem on weak convergence to stable limit processes.

THEOREM 1. *Let X be a process with independent increments satisfying $P(X(0) = 0) = 1$ and*

$$\begin{aligned} \mathbf{E}(\exp\{i\lambda(X(t) - X(s))\}) = \exp\left\{ (t-s)a_1 \int_{-\infty}^0 (e^{i\lambda u} - 1 - i\lambda u^*) d(|u|^{-\alpha}) \right. \\ \left. - (t-s)a_2 \int_0^{\infty} (e^{i\lambda u} - 1 - i\lambda u^*) d(u^{-\alpha}) \right\}, \end{aligned}$$

where $a_1, a_2 \geq 0$, $a_1 + a_2 > 0$, $0 < \alpha < 2$, $0 \leq s \leq t \leq 1$, $\lambda \in \mathbf{R}$. In order that $H_{n,h} \Rightarrow X$, it is necessary and sufficient that for any $u > 0$,

$$\sum_{a(j) < -u\beta(n), j \leq n} j^{-1} \rightarrow a_1 \theta^{-1} u^{-\alpha}, \quad \sum_{a(j) > u\beta(n), j \leq n} j^{-1} \rightarrow a_2 \theta^{-1} u^{-\alpha} \quad (7)$$

and

$$\lim_{\varepsilon \rightarrow 0} \limsup_{n \rightarrow \infty} \beta^{-2}(n) \sum_{|a(j)| < \varepsilon \beta(n), j \leq n} a(j)^2 j^{-1} = 0. \quad (8)$$

Under conditions (7) and (8), we also have $X_{n,h} \Rightarrow X$. The first result with Brownian motion as the limiting process was proved by DeLaurentis and Pittel (1985) in the case $\theta = 1$ for the function $w(\sigma)$, counting the number of cycles in (3). The case of general θ for the function $w(\sigma)$ was examined by Hansen (1990) and Donnelly *et al.* (1991). For an arbitrary additive function and $\theta > 0$, Babu and Manstavičius (1998a) obtained necessary and sufficient conditions for weak convergence to the Brownian motion. Their result generalizes the classical central limit theorem for the number of cycles (Goncharov, 1942) as well as the central limit theorem for the logarithm of the product of lengths of cycles (Erdős and Turán, 1965).

As in Probabilistic Number Theory (see Babu, 1973, Manstavičius, 1984, 1985; Kubilius, 1964), the proof of sufficiency depends on "truncated" sums, and it is

enough to establish the result for completely additive functions. The idea of the proof of necessity comes from Manstavičius (1985) and Timofeev and Usmanov (1982).

Remark. The choice of the 'time' index function $\{y(t) : 0 \leq t \leq 1\}$ makes it possible to derive the functional limit result from one-dimensional weak convergence. However, $H_{n,h}(1) \Rightarrow X(1)$ does not imply $X_{n,h}(1) \Rightarrow X(1)$. The counter example given in (Babu and Manstavičius, 1998a) illustrates this in case X is the Brownian Motion. A counter example is constructed in the next section to illustrate this fact, when $H_{n,h}(1)$ converges to a stable law.

5. COUNTER EXAMPLE

In this section, for each $0 < \alpha < 2$, an additive function h is constructed such that the distribution of $H_{n,h}(1)$ under $\nu_{n,1}$ converges weakly to a symmetric stable distribution with characteristic exponent α , while $X_{n,h}(1)$ has a different limiting distribution. The idea of construction has its origin in (Timofeev, 1985). The main construction depends on the following analytic result.

LEMMA (Manstavičius, 1996). *Let $\theta = 1$, and let f be a complex valued completely multiplicative function defined on S_n by $f_j(k) = f_j(1)^k$, where $|f_j(1)| \leq 1$, $j \geq 1$. If*

$$\sum_{j=1}^n \frac{1 - \Re f_j(1)}{j} \leq M < \infty,$$

then the mean-value of f , under $\nu_{n,1}$,

$$\begin{aligned} \frac{1}{n!} \sum_{\sigma \in S_n} f(\sigma) &= \left(1 + O\left(\frac{K}{n}\right)\right) \frac{1}{2\pi i} \int_{1-Ki}^{1+Ki} \frac{e^z}{z} \exp\left(\sum_{j=1}^n \frac{f_j(1) - 1}{j} e^{-zj/n}\right) dz \\ &+ O(K^{-(1/2)+\delta}), \end{aligned} \quad (9)$$

for $1 < K < n$ and for each $\delta \in (0, 1/2)$. The constants in the symbol O depend at most on M and δ .

THEOREM 2. *For each $0 < \alpha < 2$. There exists a sequence of numbers $\{a(j)\}$ such that the completely additive function $h(\sigma) = \sum_{j=1}^n a(j)k_j(\sigma)$ satisfies*

$$\nu_{n,1}(\beta(n)^{-1}h(\sigma) - A(n, n; h) \leq x) \rightarrow F(x) \quad (10)$$

for all $x \in \mathbb{R}$, where $\beta(n) = n^{1/\alpha}$, and F denotes the stable law with characteristic function ϕ_α given by $\phi_\alpha(s) = e^{-|s|^\alpha}$. However the distribution of the corresponding sum of independent random variables,

$$X_{n,h}(1) = n^{-1/\alpha} \sum_{j \leq n} a(j)\xi_j - A(n, n; h)$$

does not converge to F .

Proof. Define the bivariate distribution G by

$$G(x, y) = \begin{cases} F(x)y & \text{if } 0 < y < 1, \\ F(x) & \text{if } y \geq 1, \\ 0 & \text{otherwise,} \end{cases}$$

$$d(j) = \begin{cases} F^{-1}(\{j\sqrt{2}\}) & \text{if } |F^{-1}(\{j\sqrt{2}\})| \leq j^{1/\alpha}, \\ 0 & \text{otherwise,} \end{cases}$$

where $\{x\}$ denotes the fractional part of x , and

$$h(\sigma) = \sum_{j=1}^n d(j)j^{1/\alpha}k_j(\sigma).$$

Let $\mu_n(\dots) = n^{-1}\#\{1 \leq j \leq n : \dots\}$. Since $1 - F(x) + F(-x) \leq cx^{-\alpha}$, for some constant c and all $x > 1$, we have

$$\begin{aligned} \mu_n(d(j) \leq x) &= \mu_n(\{j\sqrt{2}\} \leq F(x)) + O(n^{-1/2}) \\ &\quad + O(\mu_n(\sqrt{n} < j \leq n, |F^{-1}(\{j\sqrt{2}\})| > j^{1/\alpha})) \\ &= F(x) + O(n^{-1/2}) + O(\mu_n(|F^{-1}(\{j\sqrt{2}\})| \geq n^{1/2\alpha})) \\ &= F(x) + O(n^{-1/2}) + O(F(-n^{1/2\alpha}) + 1 - F(n^{1/2\alpha})) \\ &= F(x) + O(n^{-1/2}), \end{aligned}$$

uniformly in $x \in \mathbf{R}$. The joint empirical distribution

$$\begin{aligned} F_n(x, y) &= \frac{1}{n}\#\left\{1 \leq j \leq n : d(j) \leq x, \frac{j}{n} \leq y\right\} \\ &= \frac{1}{n}\#\{1 \leq j \leq ny^* : d(j) \leq x\} \rightarrow G(x, y), \end{aligned}$$

for $x \in \mathbf{R}$ and $y > 0$. Hence $F_n \Rightarrow G$. We require the inequalities

$$|e^{ix} - 1| \leq 2|x^*| \leq 2|x^*|^\rho, \tag{11}$$

for $0 < \rho < 1$, and for $n \geq 2$,

$$\begin{aligned} \frac{1}{n} \sum_{j=1}^n |d(j)|^{\alpha/2} &= \frac{\alpha}{2} \int_0^{n^{1/\alpha}} x^{(\alpha/2)-1} \mu_n(|d(j)| \geq x) dx \\ &= \frac{\alpha}{2} \int_0^{n^{1/\alpha}} x^{(\alpha/2)-1} (1 - F(x) + F(-x)) dx + O(n^{-1/2}) \int_0^{n^{1/\alpha}} x^{(\alpha/2)-1} dx \\ &\ll \int_1^\infty x^{-(\alpha/2)-1} dx + O(1) \ll 1. \end{aligned} \tag{12}$$

For $x \in \mathbf{R}$, $0 < y \leq 1$ and complex number z , define

$$g(x, y; z) = \frac{1}{y}(e^{ixy^{1/\alpha}} - 1)e^{-zy}$$

and

$$\begin{aligned} S_n(s, z) &= \sum_{j=1}^n \frac{1}{j} (\exp\{isd(j)(j/n)^{1/\alpha}\} - 1) e^{-zj/n} \\ &= \int g(sx, y; z) dF_n(x, y). \end{aligned}$$

Note that for each fixed z, s , $g(sx, y; z)$ as a function of x, y is continuous and hence $F_n g^{-1} \Rightarrow G g^{-1}$. If s is real and $1 < \beta < (2/\alpha) \wedge (3/2)$, then by applying (11) with $\rho = \alpha/(2\beta)$ we get, uniformly in $\Re z \geq 0$, that

$$\begin{aligned} |S_n(s, z)|^\beta &\leq \int |g(sx, y; z)|^\beta dF_n(x, y) = \frac{1}{n} \sum_{j=1}^n |g(sd(j), (j/n); z)|^\beta \\ &\ll \frac{1}{n} \sum_{j=1}^n \left| \frac{1}{j} (sd(j)(j/n)^{1/\alpha})^* \right|^\beta \ll \frac{1}{n} \sum_{j=1}^n |d(j)|^{\alpha/2} (j/n)^{(1/2)-\beta} \\ &\ll 1. \end{aligned} \quad (13)$$

Summation by parts and (12) are used to obtain the last inequality above. For $\Re z \geq 0$ and s real, the inequalities (13) imply that $\sup_n \int |u|^\beta dF_n g^{-1} < \infty$, for some $\beta > 1$, and hence the identity map is uniformly integrable under the sequence of measures $F_n g^{-1}$. Thus for $\Re z \geq 0$, real s and for some $\beta > 1$,

$$\int |g(sx, y; z)| dG(x, y) < \infty, \quad \int |g(sx, y; z)|^\beta dG(x, y) < \infty, \quad (14)$$

and

$$S_n(s, z) = \int g(sx, y; z) dF_n(x, y) \longrightarrow \int g(sx, y; z) dG(x, y). \quad (15)$$

Similarly,

$$A(n, n; h) = \int \frac{1}{y} (xy^{1/\alpha})^* dF_n(x, y) \longrightarrow \int \frac{1}{y} (xy^{1/\alpha})^* dG(x, y) = 0. \quad (16)$$

The last equality in (16) follows as $F(x) = 1 - F(-x)$ for all real x . Hence $A(n, n; h) \rightarrow 0$. By (13), $S_n(s, z)$ is bounded uniformly in $\Re z \geq 0$. Hence we have by the dominated convergence theorem, that for any $K > 0$,

$$\begin{aligned} &\frac{1}{2\pi i} \int_{1-iK}^{1+iK} \frac{e^z}{z} (\exp(S_n(s, z))) dz \\ &\longrightarrow \frac{1}{2\pi i} \int_{1-iK}^{1+iK} \frac{e^z}{z} \left(\exp\left(\int g(sx, y; z) dG(x, y) \right) \right) dz. \end{aligned} \quad (17)$$

Note that by (14) and Fubini's theorem, we have for any z with non-negative real part,

$$\begin{aligned} \int g(sx, y; z) dG(x, y) &= \int_0^1 \int_{\mathbf{R}} \frac{1}{y} (e^{isxy^{1/\alpha}} - 1) e^{-zy} dy dF(x) \\ &= \int_0^1 \frac{1}{y} e^{-zy} (e^{y|s|^\alpha} - 1) dy = \log \left(\frac{z}{z + |s|^\alpha} \right) - \int_{z+|s|^\alpha}^z \frac{e^{-w}}{w} dw. \end{aligned}$$

Consequently for $K > 1$,

$$\begin{aligned} &\frac{1}{2\pi i} \int_{1-iK}^{1+iK} \frac{e^z}{z} \exp \left(\int g(sx, y; z) dG(x, y) \right) dz \\ &= \frac{1}{2\pi i} \int_{1-iK}^{1+iK} \frac{e^z}{z + |s|^\alpha} \exp \left(\int_0^{|s|^\alpha} \frac{e^{-\omega-z}}{\omega + z} d\omega \right) dz \\ &= \frac{1}{2\pi i} \int_{1-iK}^{1+iK} \frac{e^z}{z + |s|^\alpha} dz + \frac{1}{2\pi i} \int_{1-iK}^{1+iK} \frac{e^z}{z + |s|^\alpha} \\ &\quad \times \left(\exp \left(\int_z^{z+|s|^\alpha} \frac{e^{-w}}{w} dw \right) - 1 \right) dz + O(K^{-1/2+\delta}). \end{aligned} \quad (18)$$

The integrals occurring in (18) have been investigated in (Timofeev, 1985). The last but one integral in (18) satisfies

$$\frac{1}{2\pi i} \int_{1-iK}^{1+iK} \frac{e^z}{z + |s|^\alpha} dz = \exp(-|s|^\alpha) + O(K^{-1}). \quad (19)$$

The bound $O((\log K)/K)$ for the last integral in (18) is obtained by changing the range of integration $[1 - iK, 1 + iK]$ using the contour $\{z : \Im z = \pm K, 1 \leq \Re z \leq K\}$, $\{z : \Re z = K, |\Im z| \leq K\}$, and using the estimate

$$\exp \left(\int_z^{z+|s|^\alpha} \frac{e^{-w}}{w} dw \right) - 1 = O \left(\frac{e^{-\Re z}}{|z|} \right)$$

on the contour. Thus we obtain by (17), (18), (19) and the lemma, that $v_{n,1}(h(\sigma) < xn^{1/\alpha}) \rightarrow F(x)$, uniformly in $x \in \mathbf{R}$ as $n \rightarrow \infty$. This implies (10) as $A(n, n; h) \rightarrow 0$ by (16).

On the other hand the characteristic function of $n^{-1/\alpha} \sum_{j \leq n} h_j(1)\xi_j$ is given by

$$\begin{aligned} \psi_n(s) &= \exp(S_n(s, 0)) \longrightarrow \exp\left(\int g(sx, y; 0) dG(x, y)\right) \\ &= \exp\left(\int_0^1 \frac{1}{y} (e^{-y|s|^\alpha} - 1) dy\right) = \exp\left(\int_0^1 \frac{1}{y} (\phi_\alpha(sy^{1/\alpha}) - 1) dy\right) \neq \phi_\alpha(s). \end{aligned} \quad (20)$$

Since $A(n, n; h) \rightarrow 0$, $X_{n,h}(1)$ also converges weakly to the distribution with the characteristic function given by (20). This completes the proof.

REFERENCES

- Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.* **2**, 1152–1174.
- Arratia, R., Barbour, A. D. and Tavaré, S. (1992). Poisson process approximations for the Ewens sampling formula. *Ann. Appl. Probab.* **2**, 519–535.
- Babu, G. J. (1973). A note on the invariance principle for additive functions. *Sankhyā A* **35**, 307–310.
- Babu, G. J. and Manstavičius, E. (1998a). Brownian motion for random permutations. Submitted for publication.
- Babu, G. J. and Manstavičius, E. (1998b). Infinitely divisible limit processes for the Ewens sampling formula. Submitted for publication.
- Billingsley, P. (1968). *Convergence of Probability Measures*. Wiley, New York.
- DeLaurentis, J. M. and Pittel, B. G. (1985). Random permutations and the Brownian motion. *Pacific J. Math.* **119**, 287–301.
- Donnelly, P., Kurtz, T. G., and Tavaré, S. (1991). On the functional central limit theorem for the Ewens Sampling Formula. *Ann. Appl. Probab.* **1**, 539–545.
- Erdős, P. and Turán, P. (1965). On some problems of a statistical group theory I. *Z. Wahrsch. Verw. Gebiete* **4**, 175–186.
- Erdős, P. and Turán, P. (1967). On some problems of a statistical group theory III. *Acta Math. Acad. Sci. Hungar.* **18**, 309–320.
- Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theor. Pop. Biol.* **3**, 87–112.
- Goncharov, V. L. (1942). On the distribution of cycles in permutations. *Dokl. Acad. Nauk SSSR* **35**, 299–301 (in Russian).
- Hansen, J. C. (1990). A functional central limit theorem for the Ewens Sampling Formula. *J. Appl. Probab.* **27**, 28–43.
- Karlin, S. and McGregor, J. L. (1972). Addendum to a paper of W. Ewens. *Theor. Pop. Biol.* **3**, 112–116.
- Kingman, J. F. C. (1980). *Mathematics of Genetic Diversity*. SIAM, Philadelphia, PA.
- Kubilius, J. (1964). *Probabilistic Methods in the Theory of Numbers*. Amer. Math. Soc. Transl. **11**, Providence, RI.
- Manstavičius, E. (1984). Arithmetic simulation of stochastic processes. *Lith. Math. J.* **24**, 276–285.
- Manstavičius, E. (1985). Additive functions and stochastic processes. *Lith. Math. J.* **25**, 52–61.
- Manstavičius, E. (1996). Additive and multiplicative functions on random permutations. *Lith. Math. J.* **36**, 400–408.
- Timofeev, N. M. and Usmanov, H. H. (1982). Arithmetic simulation of the Brownian motion. *Dokl. Acad. Nauk Tadzh. SSR* **25**, 207–211 (in Russian).
- Timofeev, N. M. (1985). Stable limit laws for additive arithmetic functions. *Mat. Zametki* **37**, 465–473 (in Russian).