

BROWNIAN MOTION FOR RANDOM PERMUTATIONS

By G.J. BABU*

The Pennsylvania State University, University Park

and

E. MANSTAVIČIUS**

Vilnius University, Vilnius

SUMMARY. A family of measures, on the set of partitions of an integer, known as the *Ewens sampling formula* arises in population genetics. Mixtures of these measures also have applications in Bayesian statistics. Using methods from probabilistic number theory, a functional limit theorem in $\mathbf{C}[\mathbf{0}, \mathbf{1}]$ is established for a partial sum process based on these measures. The results can be used to develop statistical methods to test the validity of certain genetic models. It is interesting to note that a Lindeberg type condition is necessary for the dependent process to converge to the Brownian Motion, while it is not the case for the convergence of the one dimensional distributions. An example to illustrate this phenomenon is constructed.

1. Introduction

In the last few decades, *mathematical population geneticists* have been exploring the mechanisms that maintain diversity in a population. Natural selection due to interaction with the environment and mutation are some of the factors to be taken into account. Some geneticists believe that much of genetic diversity occurs mainly due to mutation and random fluctuations that are inherent in the reproductive process. They view that the effect of selection has been exaggerated. Ewens (1972) established an approximation to the sampling distribution of a sample of n genes from a population that has evolved over several generations, by a partly heuristic argument. The derivation ignores the

Paper received. March 1999.

AMS (1991) subject classification. Primary 60F17; secondary 60C05, 11K65.

Key words and phrases. Ewens sampling formula, random partitions, functional limit theorem, additive functions, multiplicative functions.

* Research supported in part by NSA grant MDA904-97-1-0023 and NSF grant DMS-9626189.

** Research supported in part by National Research Council's 1997-99 Twinning fellowship. E. Manstavičius acknowledges the hospitality of the Pennsylvania State University, where a part of this manuscript was prepared.

selective effects and assumes that there is no meaningful way of labelling the alleles. In this case the *allelic partition* $\bar{k} = (k_1, \dots, k_n)$, where k_j denotes the number of alleles represented j times in the sample, $j = 1, \dots, n$, contains all the information available in a sample of n genes.

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

$$\nu_{n,\theta}(k_1, \dots, k_n) := \frac{n!}{\theta^{(n)}} \prod_{j=1}^n \left(\frac{\theta}{j}\right)^{k_j} \frac{1}{k_j!}, \quad \dots (1.1)$$

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

$$1k_1 + \cdots + nk_n = n. \quad \dots (1.2)$$

The vector \bar{k} represents a partition of the integer n . The derivation of equation (1.1) was made rigorous by Karlin & McGregor (1972). Kingman (1980) argues that several different approaches lead to (1.1), “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 non-recurrent, so that every mutant allele is a completely novel one.”

The Ewens’ formula can be used to test if these assumptions are consistent with data, and to estimate the parameter θ . The statistics that are useful in this connection will 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.

Antoniak (1974) considers mixtures of measures described by the Ewens’ formula in Bayesian non-parametric statistics. A review of the Ewens Sampling Formula and its applications can be found in Chapter 41 of Johnson *et al.* (1997) and Kotz *et al.* (1998, pp. 230-234).

A functional limit theorem for a partial sum process $\sum_{j \leq y(t)} h_j(k_j), 0 \leq t \leq 1$ is established in this paper, where y is a function on the unit interval. The derivations involve concepts and ideas from probabilistic number theory.

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 can be seen by partitioning S_n into equivalence classes of conjugate elements and identifying a conjugate class with a vector representing numbers of cycles of various lengths. More precisely, let $\sigma \in S_n$ be an arbitrary permutation and

$$\sigma = \kappa_1 \cdots \kappa_w, \quad \dots (1.3)$$

be its unique representation (up to the order) as a product of independent cycles κ , where $w = w(\sigma)$ denotes the number of cycles. According to the general theory, the elements in a conjugate class will all have the same number k_j of cycles of length j for all $1 \leq j \leq n$. Hence the conjugate class containing an element $\sigma \in S_n$, expressed as in (1.3), can be identified 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 a function on S_n . This motivates a study of the distribution of values of a function on S_n .

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

$$h(\sigma) = \sum_{j=1}^n h_j(k_j(\sigma)) \quad \dots (1.4)$$

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 \mathbf{G} . If $h_j(k) = kh_j(1) =: ka(j)$ for each $1 \leq j \leq n$ and $k \geq 0$, then h is called *completely additive function* (linear statistics). The number of cycles $w(\sigma)$ in (1.3) is a typical example of such functions.

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 (1.1) is a probability measure on \mathcal{F} . The uniform distribution (the Haar measure) on S_n induces the measure $\nu_n = \nu_{n,1}$ on the space of conjugate classes. This is the Ewens' formula when $\theta = 1$. The results of this paper are based on the relation

$$\nu_{n,\theta}(\bar{k}) = P(\xi_1 = k_1, \dots, \xi_n = k_n \mid 1\xi_1 + \dots + n\xi_n = n), \quad \dots (1.5)$$

where ξ_j , $1 \leq j \leq n$ are independent Poisson random variables, satisfying $\mathbf{E}\xi_j = \theta/j$ (see, for instance, Arratia *et al.*, 1992).

We establish general invariance principle for additive functions $h(\sigma)$ given by (1.4). Define

$$H_n := H_n(\sigma, t) = \frac{1}{B(n)} \left(\sum_{j \leq y(t)} h_j(k_j(\sigma)) - A(y(t)) \right),$$

where

$$A(u) := \sum_{j \leq u} \frac{a(j)}{j} \theta, \quad B^2(u) := \sum_{j \leq u} \frac{a(j)^2}{j} \theta,$$

$h_j(1) =: a(j)$, and

$$y(t) := y_n(t) = \max\{u \leq n : B^2(u) \leq tB^2(n)\}, \quad t \in [0, 1].$$

We consider the weak convergence (denoted by \Rightarrow) of the process H_n . H_n is a random element in the space $\mathbf{D}[0, 1]$ equipped with the supremum norm. Here and in what follows the limits are taken with respect to $n \rightarrow \infty$. Observe that in the case of the limiting Wiener measure W , it is natural to use the uniform metric (see Billingsley, 1968). Of course, one could also examine a linearized version of the process H_n and deal only with elements of the space $\mathbf{C}[0, 1]$. The main result of the present paper is the following theorem.

THEOREM 1. *Let $h(\sigma)$ be a real additive function, $h_j(1) = a(j)$, $B(n) \rightarrow \infty$. For the weak convergence*

$$\nu_{n,\theta} \cdot H_n^{-1} \Rightarrow W \quad \dots (1.6)$$

to hold it is necessary and sufficient that, for each $\varepsilon > 0$,

$$\Lambda_n(\varepsilon) := \frac{1}{B^2(n)} \sum_{\substack{j=1 \\ |a(j)| \geq \varepsilon B(n)}}^n \frac{a(j)^2}{j} = o(1). \quad \dots (1.7)$$

The first functional limit theorem in the case $\theta = 1$ for the function $w(\sigma)$, counting the number of cycles in (1.3), was proved by DeLaurentis and Pittel (1985). The case of general θ for the function $w(\sigma)$ was examined by Hansen (1990), Donnelly *et al.* (1991), and Arratia & Tavaré (1992).

In proving the sufficiency of (1.7) in the general case, as in probabilistic number theory Babu (1973), Kubilius (1964), Manstavičius (1984 and 1985), we use an appropriate total variation distance between the truncated sums of dependent and the corresponding independent Poisson variables. This idea has also been noted in Theorem 6 of Arratia & Tavaré (1992). To show that “truncated parts” of additive functions are negligible, we use a generalization of the inequalities obtained in Manstavičius (1998). This provides a more general approach than that used by Arratia & Tavaré (1992, Theorem 5) in the proof of Hansen’s theorem. The idea of our proof of necessity comes from Manstavičius (1985) and Timofeev & Usmanov (1982). Some of the techniques are borrowed from Probabilistic Number Theory.

The relation (1.6) implies the convergence

$$\nu_{n,\theta}(h(\sigma) - A(n) < xB(n)) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-u^2/2} du + o(1). \quad \dots (1.8)$$

Thus Theorem 1 generalizes the classical result of Goncharov (1942) as well as the central limit theorem for the logarithm of the product of lengths of cycles of a random σ (see Erdős & Turán, 1965). It is easy to check that the independent

random variables $\xi_j/B(n)$ (approximating $k_j(\sigma)/B(n)$ for small j) are infinitesimal. Therefore one could expect that the Lindeberg type condition (1.7) is necessary for the relation (1.8). But it is not the case, since the dependence of random classes of permutations with large cycle lengths plays a substantial role. In the Section 4, we include a counter example to show that (1.7) is not necessary for the one-dimensional weak convergence result for $H_n(\cdot, 1)$, unlike the Lindeberg-Feller theorem for the corresponding independent random variables.

2. Auxiliary Results

We shall show later that our problem easily reduces to that for completely additive functions. We define

$$\hat{H}_n := \hat{H}_n(\sigma, t) := \frac{1}{B(n)} \left(\sum_{j \leq y(t)} a(j)k_j(\sigma) - A(y(t)) \right) \quad \dots (2.1)$$

and

$$X_n := X_n(t) = \frac{1}{B(n)} \left(\sum_{j \leq y(t)} a(j)\xi_j - A(y(t)) \right).$$

Let $\hat{H}_n^r := \hat{H}_n^r(\sigma, t)$ and $X_n^r := X_n^r(t)$ be the processes obtained from \hat{H}_n and X_n respectively by substituting $y(t) \wedge r$ for $y(t)$. Here $2 \leq r \leq n$ and $a \wedge b := \min\{a, b\}$. Let $\|\cdot\|$ denote the total variance of signed measures. We now present two main lemmas. The first lemma, which is needed in the proof of sufficiency, is an extension of Corollary 5.1 in Manstavičius (1998). Let for brevity, $\mathcal{L}(I)$ be the linear space of real functions $g(\cdot)$ on $I \subset \mathbf{R}$ with $\sup_{t \in I} |g(t)| < \infty$.

LEMMA 1. *Let $h(\sigma, t)$, $t \in I \subset \mathbf{R}$, be a set of real valued additive functions defined by (1.4) via $h_j(k, \cdot) \in \mathcal{L}(I)$, where $k \geq 0$, $h_j(0, t) = 0$, for $j \leq n$, $t \in I$, and $\Xi_n(t) = h_1(\xi_1, t) + \dots + h_n(\xi_n, t)$. Then*

$$\nu_{n,\theta} \left(\sup_{t \in I} |h(\sigma, t) - a(t)| \geq u \right) \leq C(\theta) \left(P^{\theta \wedge 1} \left(\sup_{t \in I} |\Xi_n(t) - a(t)| \geq u/3 \right) + n^{-\theta} \right).$$

Here $a(\cdot) \in \mathcal{L}(I)$ and $u \geq 0$ are arbitrary and $C(\theta)$ is a positive constant depending only on θ .

PROOF. For $\theta \geq 1$, the lemma follows from Theorem 4 of Manstavičius (1998) as in his Corollary 5.1. For $\theta < 1$, the lemma follows from Lemma A (see the Appendix), which is a technical result of independent interest on lattice valued random variables.

For the proof of necessity of (1.7) we need an estimate of the mean values of multiplicative functions defined on permutations having only long cycles.

LEMMA 2. Let $f : S_n \rightarrow \mathbf{C}$ be a completely multiplicative function defined by $f_j(k) = b(j)^k$ where $b(j) = 1$ for all but $j \in J \subset (n/2, n]$. Then

$$M_n(f) := \frac{n!}{\theta_{(n)}} \sum_{\bar{k}} \prod_{j=1}^n \left(\frac{\theta}{j}\right)^{k_j} \frac{f_j(k_j)}{k_j!} = 1 + \theta \sum_{j \in J} \frac{b(j) - 1}{j} \frac{n!}{\theta_{(n)}} \frac{\theta_{(n-j)}}{(n-j)!}.$$

Moreover, if $|b(j)| \leq 1$ and $J \subset ((1 - \delta)n, n]$ for $0 < \delta < \delta(\theta) < 1/2$, then

$$|M_n(f)| > c(\theta) > 0$$

provided $\delta(\theta)$ is sufficiently small and n is sufficiently large, $n > n(\theta)$.

PROOF. Observe that, if $k_j \geq 1$ for some $j \in J$, then the $1k_1 + \dots + nk_n = n$ implies $k_j = 1$ and $k_l = 0$ for the remaining $l \neq j$ and $l \in J$. Hence

$$\begin{aligned} M_n(f) &= \frac{n!}{\theta_{(n)}} \sum_{\substack{\bar{k} \\ k_l = 0 \forall l \in J}} \prod_{l=1}^n \left(\frac{\theta}{l}\right)^{k_l} \frac{1}{k_l!} + \frac{n!}{\theta_{(n)}} \sum_{j \in J} b(j) \sum_{\substack{\bar{k} \\ k_j = 1}} \prod_{l=1}^n \left(\frac{\theta}{l}\right)^{k_l} \frac{1}{k_l!} \\ &= 1 + \frac{n!}{\theta_{(n)}} \sum_{j \in J} (b(j) - 1) \sum_{\substack{\bar{k} \\ k_j = 1}} \prod_{l=1}^n \left(\frac{\theta}{l}\right)^{k_l} \frac{1}{k_l!} \\ &= 1 + \sum_{j \in J} (b(j) - 1) \left(1 - \frac{n!}{\theta_{(n)}} \sum_{\substack{\bar{k} \\ k_j = 0}} \prod_{l=1}^n \left(\frac{\theta}{l}\right)^{k_l} \frac{1}{k_l!}\right) \\ &= 1 + \sum_{j \in J} (b(j) - 1) \left(1 - \frac{n!}{\theta_{(n)}} d_j(n)\right), \end{aligned} \tag{2.2}$$

where

$$d_j(n) = \sum_{\substack{\bar{k} \\ k_j = 0}} \prod_{l=1}^n \left(\frac{\theta}{l}\right)^{k_l} \frac{1}{k_l!}.$$

To calculate the sum $d_j(n)$, we compare two Taylor series expansions of the function

$$\begin{aligned} (1 - z)^{-\theta} \exp\{-\theta z^j / j\} &= \sum_{n \geq 0} d_j(n) z^n \\ &= \sum_{\ell \geq 0} q_\ell z^\ell \sum_{m \geq 0} \frac{\theta_{(m)}}{m!} z^m. \end{aligned}$$

Here $q_\ell = 0$ if j does not divide ℓ and

$$q_{jk} = \left(\frac{-\theta}{j}\right)^k \frac{1}{k!}.$$

Hence

$$d_j(n) = \sum_{\substack{k, m \geq 0 \\ jk + m = n}} q_{jk} \frac{\theta_{(m)}}{m!}.$$

If $j \in J$, then there are only two nonzero summands in the last sum. Thus

$$d_j(n) = \frac{\theta_{(n)}}{n!} - \frac{\theta}{j} \frac{\theta_{(n-j)}}{(n-j)!}.$$

The proof of the first part of the lemma is completed by substituting this into (2.2).

If $|b(j)| \leq 1$, $\theta \geq 1$, $\delta < 1/(4\theta)$ and n is sufficiently large, then trivially

$$|M_n(f)| \geq 1 - 2\theta(-\log(1 - \delta) + o(1)) \geq 1/4.$$

If $\theta < 1$, then using the relation $\theta_{(n)}/n! \sim n^{\theta-1}/\Gamma(\theta)$, we obtain the estimate

$$\frac{n!}{\theta_{(n)}} \frac{\theta_{(n-j)}}{(n-j)!} \leq C(\theta) \left(\frac{n}{n-j+1} \right)^{1-\theta},$$

where Γ denotes the Euler gamma-function. We have for $n\delta \geq 1$,

$$|M_n(f)| \geq 1 - 4C(\theta)n^{-\theta} \sum_{m \leq 2\delta n} m^{\theta-1} \geq 1 - 8C(\theta)\theta^{-1}\delta^\theta + o_\theta(1).$$

The second part now follows, if $\delta(\theta)$ is chosen to be sufficiently small number. This completes the proof of Lemma 2.

3. Proof of Theorem 1

Sufficiency. Recall the definition of the function $\widehat{H}_n(\sigma, t)$ from (2.1). We first note that, without loss of generality, one may restrict to the class of completely additive functions. Indeed, if $\delta > 0$ and $K > 2$ are arbitrary, then by Lemma 1 and since $B(n) \rightarrow \infty$, we have

$$\begin{aligned} \nu_{n,\theta}(\delta) &= \nu_{n,\theta} \left(\sup_t |H_n(\sigma, t) - \widehat{H}_n(\sigma, t)| > \delta \right) \\ &\ll P^{\theta \wedge 1}(\exists j \leq K : \xi_j \geq K) + P^{\theta \wedge 1}(\exists j \geq K : \xi_j \geq 2) \\ &\quad + P^{\theta \wedge 1} \left(\sum_{j \leq K} (|h_j(\xi_j)| + |a(j)|\xi_j) \geq \delta B(n)/3, \xi_j \leq K \forall j \leq K \right) + o(1) \\ &\ll \left(\sum_{j \leq K} \sum_{k \geq K} e^{-\theta/j} \frac{\theta^k}{j^k k!} \right)^{\theta \wedge 1} + \left(\sum_{j \geq K} \sum_{k \geq 2} e^{-\theta/j} \frac{\theta^k}{j^k k!} \right)^{\theta \wedge 1} + o_K(1) \\ &\ll \left(\sum_{j \leq K} \frac{\theta^K}{j^K K!} \right)^{\theta \wedge 1} + \left(\sum_{j \geq K} \frac{\theta^2}{j^2} \right)^{\theta \wedge 1} + o_K(1) \\ &\ll K^{-\theta \wedge 1} + o_K(1). \end{aligned}$$

Hence $\nu_{n,\theta}(\delta) = o(1)$ for arbitrary $\delta > 0$, and the measures $\nu_{n,\theta} \cdot H_n^{-1}$ and $\nu_{n,\theta} \cdot \widehat{H}_n^{-1}$ can only converge simultaneously. Thus, from now on we analyze the processes \widehat{H}_n and \widehat{H}_n^r omitting the “hats”.

Further, note that if condition (1.7) holds, then for each $0 < u < 1$,

$$B^2(n) - B^2(un) \leq \varepsilon^2 B^2(n) \sum_{un < j \leq n} \frac{1}{j} + B^2(n)\Lambda_n(\varepsilon) \leq \left(\varepsilon^2 \log \frac{1}{u} + o(1) \right) B^2(n).$$

Hence $B(un) \sim B(n)$, in other words, $B(n)$ is slowly varying in the sense of Karamata. Moreover, there exists $r = u_n n \rightarrow \infty$ with $u_n \rightarrow 0$ such that $B(r) \sim B(n)$. We use this r in the definitions above of H_n^r and X_n^r .

By the weak invariance principle, it follows from (1.7) (see, for instance, Manstavičius, 1985) that $P \cdot X_n^{-1} \Rightarrow W$. Let $t_n = \sup\{t : y(t) \leq r\}$. Since for each $\delta > 0$,

$$\begin{aligned} P_n(\delta) &:= P\left(\sup_t |X_n(t) - X_n^r(t)| \geq \delta\right) \\ &\leq P\left(\sup_{t_n \leq t \leq 1} \left| \sum_{r < j \leq y(t)} a(j) \left(\xi_j - \frac{\theta}{j}\right) \right| > \delta B(n)\right) \\ &\leq P\left(\sum_{r \leq j \leq n} |a(j)| \left|\xi_j - \frac{\theta}{j}\right| \geq \delta B(n)\right) \\ &\leq \frac{B^2(n) - B^2(r)}{\delta^2 B^2(n)} = o(1), \end{aligned} \tag{3.1}$$

we also have $P \cdot X_n^{r^{-1}} \Rightarrow W$. We now use the estimate of Arratia *et al.*, (1992, Theorem 3),

$$\|\nu_{n,\theta} \cdot \widehat{H}_n^{j^{-1}} - P \cdot X_n^{j^{-1}}\| \leq \frac{j\theta}{n}(\theta + n(n-j)^{-1}), \quad 2 \leq j < n, \tag{3.2}$$

on the total variation distance between the truncated sums of dependent and the corresponding independent Poisson random variables. Hence by (3.2), we have $\nu_{n,\theta} \cdot H_n^{r^{-1}} \Rightarrow W$.

Finally, using Lemma 1 and (3.1), we obtain

$$\nu_{n,\theta} \left(\sup_t |H_n(\sigma, t) - H_n^r(\sigma, t)| \geq \delta \right) \leq C(\theta) P_n^{1 \wedge \theta}(\delta/3) + o(1) = o(1),$$

for each $\delta > 0$. Hence we see that $\nu_{n,\theta} \cdot H_n^{-1}$ also converges weakly to W .

Necessity. Let $\nu_{n,\theta} \cdot H_n^{-1} \Rightarrow W$. Then for each $0 \leq t < 1$, the difference $H_n(\sigma, 1) - H_n(\sigma, t)$ converges weakly to the normal distribution with zero mean and variance $1 - t$. Let $\varphi_n(u, t)$, $u \in \mathbf{R}$, denote the characteristic function of $H_n(\sigma, 1) - H_n(\sigma, t)$. Define $b(j) = \exp\{iua(j)/B(n)\}$, $u \in \mathbf{R}$, if $y(t) < j \leq n$

and $b(j) = 1$ elsewhere. For the completely multiplicative function f defined via $f_j(1) = b(j)$, we have

$$|\varphi_n(u, t)| = |M_n(f)| \leq e^{-u^2(1-t)/2} + o(1) \quad \dots (3.3)$$

for $u \in \mathbf{R}$ and $0 < t < 1$. For t close to 1, we will apply Lemma 2. Let $0 < \delta < \delta(\theta) < 1/2$ and $\tau_n = \sup\{t : y(t) \leq (1 - \delta)n\}$. Observe that $\tau_n \rightarrow 1$. Indeed, if $\tau_n \rightarrow t_0 < t_1 < 1$ for some subsequence $n := n' \rightarrow \infty$, then $y(t_1) \geq (1 - \delta)n$ for n sufficiently large. Now Lemma 2 yields the estimate $|\varphi_n(u, t_1)| > c(\theta) > 0$ uniformly in $u \in \mathbf{R}$, contradicting to (3.3). Thus from the definitions of $y(t)$ and the sequence τ_n , it follows that

$$1 + o(1) \leq \tau_n \leq \frac{B^2(y(\tau_n) + 1)}{B^2(n)} \leq \frac{B^2((1 - \delta)n + 1)}{B^2(n)} \leq 1.$$

Hence $B(un) \sim B(n)$ for each $u \in [(1 - (\delta/2))n, n]$ and some $\delta > 0$. Substituting $(1 - (\delta/2))n$ for n repeatedly, we deduce the existence of $r = r(n) \rightarrow \infty$ such that $r = o(n)$ and $B(r) \sim B(n)$. Now repeating the arguments of the proof of the sufficiency part we obtain

$$\nu_{n,\theta}(H_n^r(\sigma, 1) < x) = \Phi(x) + o(1),$$

where Φ denotes the standard normal distribution function. This together with (3.2) leads to the central limit theorem

$$P(X_n(1) < x) = \Phi(x) + o(1).$$

Since $B(n) \rightarrow \infty$, the random variables $\{\xi_j/B(n), 1 \leq j \leq n\}$ form an *infinitesimal* array, and hence the necessity part of (1.7) follows from the Lindeberg-Feller theorem. This completes the proof of Theorem 1.

4. Counter Example to the One-Dimensional Case

In this section we present an example of an additive function having the standard normal law as its limiting distribution, but for which condition (1.7) is violated. We construct it with much larger normalizing sequence $B(n)$ than is allowed by (1.7). The idea has its origin in Timofeev (1985).

THEOREM 2. *Let $\theta = 1$. There exists a completely additive function $h(\sigma)$, $h_j(1) = a(j)$, with $A(n) = o(\sqrt{n})$ and $B(n) = \sqrt{n}(1 + o(1))$ such that (1.8) holds.*

Violation of condition (1.7) for $0 < \varepsilon < 1/2$ can be seen from

$$\Lambda_n(\varepsilon) \geq \frac{1}{n} (B^2(n) - B^2(n/2)) - \varepsilon^2 \log 2 + o(1) \geq \frac{1}{2} - \varepsilon^2 \log 2 + o(1) \geq \frac{1}{4} + o(1).$$

The main construction depends on the following analytic result.

LEMMA 3 *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 ,

$$\frac{1}{n!} \sum_{\sigma \in S_n} f(\sigma) = \left(1 + O\left(\frac{K}{n}\right)\right) \frac{1}{2\pi i} \int_{1-Ki}^{1+iK} \frac{e^z}{z} \exp\left\{\sum_{j=1}^n \frac{f_j(1) - 1}{j} e^{-zj/n}\right\} dz + O(K^{-1/2+\delta}), \dots(4.1)$$

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 .

PROOF. See Manstavičius (1996).

PROOF OF THEOREM 2. Let $\mu_n(\dots) := n^{-1} \#\{j \leq n, \dots\}$ and $\Phi^{-1}(x)$ denote the inverse function of $\Phi(x)$. We start with the following well known result on uniform convergence. By denoting the fractional part of $j\gamma$ by γ_j , we have

$$\mu_n(\gamma_j < \Phi(x)) = \Phi(x) + O(n^{-1/2}) \dots(4.2)$$

uniformly in $x \in \mathbf{R}$, for some irrational number γ (say, quadratic irrational such as $\sqrt{2}$). Define

$$d(j) = \begin{cases} \Phi^{-1}(\gamma_j) & \text{if } |\Phi^{-1}(\gamma_j)| \leq \log j, \\ 0 & \text{otherwise.} \end{cases}$$

It follows from (4.2) that

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

uniformly in $x \in \mathbf{R}$. Hence, for $n \geq 2$ and $s = 1, 2$,

$$\begin{aligned} \frac{1}{n} \sum_{j=1}^n |d(j)|^s &= s \int_0^{\log n} x^{s-1} \mu_n(|d(j)| \geq x) dx \\ &= \int_{\mathbf{R}} |x|^s d\Phi(x) + O\left(\frac{\log^2 n}{\sqrt{n}}\right). \end{aligned} \dots(4.3)$$

Define $a(j) = d(j)\sqrt{j}$ and

$$S_n(t, z) := \sum_{j=1}^n \frac{\exp\{ita(j)/\sqrt{n}\} - 1}{j} e^{-zj/n},$$

where $t \in \mathbf{R}$, $|t| \leq T$, $z = 1 + i\tau$, $|\tau| \leq K$, and $T, K, \geq 1$ are arbitrary. By (4.3), the summands in $S_n(t, z)$ when $j \leq \delta n$, $0 < \delta < 1/2$, contribute the error

$$O\left(\sum_{j \leq \delta n} \frac{|d(j)|}{\sqrt{jn}}\right) = O\left(\frac{\sqrt{\delta}}{\delta n} \sum_{j \leq \delta n} |d(j)| + \frac{1}{\sqrt{n}} \int_0^{\delta n} \frac{1}{u} \sum_{j \leq u} |d(j)| \frac{du}{\sqrt{u}}\right) = O(\sqrt{\delta}). \tag{4.4}$$

Further, the summation by parts yields that the sum of reciprocals of j , $\delta n < j \leq n$, with $\gamma_j < \delta$ or $\gamma_j > 1 - \delta$ is $O(\delta \log 1/\delta)$. This estimate and (4.4) yields that

$$\begin{aligned} S_n(t, z) &= \frac{1}{n} \sum_{j=1}^n g\left(\gamma_j, \frac{j}{n}\right) I_{\Delta}\left(\gamma_j, \frac{j}{n}\right) + O(\sqrt{\delta}) \\ &= \int_{\Delta} \int g(x, y) dx dy + O(\sqrt{\delta}) + o(1), \end{aligned} \tag{4.5}$$

where I_{Δ} denote the indicator function of the set $\Delta := [\delta, 1 - \delta] \times [\delta, 1]$ and for $(x, y) \in (0, 1) \times (0, 1]$, and

$$g(x, y) := \left(e^{it\Phi^{-1}(x)y^{1/2}} - 1 \right) \frac{e^{-zy}}{y}.$$

The last approximation of the sum by the integral could be verified at first for simple (step-wise) functions, then using a uniform approximation of $g(x, y)$ by the simple functions on Δ . The extension of the integral over the set $(0, 1) \times (0, 1]$ yields the error bound

$$O\left(\int_0^{\delta} \int_0^1 |\Phi^{-1}(x)|y^{-1/2} dx dy\right) = O\left(\sqrt{\delta} \int_{\mathbf{R}} |u| d\Phi(u)\right) = O(\sqrt{\delta})$$

for arbitrary $\delta > 0$. Hence by (4.5) and by Fubini's theorem we obtain

$$\begin{aligned} S_n(t, z) &= \int_0^1 \int_0^1 g(x, y) dx dy + o(1) \\ &= \int_0^1 \frac{e^{-zy}}{y} \left(\int_{\mathbf{R}} (e^{itu\sqrt{y}} - 1) d\Phi(u) \right) dy + o(1) \\ &= \int_0^1 \left(e^{-t^2 y/2} - 1 \right) \frac{e^{-zy}}{y} dy + o(1) \\ &= \int_0^1 \int_{z+t^2/2}^z e^{-wy} dw dy + o(1) \\ &= \log \frac{z}{z+t^2/2} - \int_{z+t^2/2}^z \frac{e^{-w}}{w} dw + o(1) \end{aligned} \tag{4.6}$$

uniformly in t , $|t| \leq T$, and τ , $|\tau| \leq K$.

Define the completely additive function $h(\sigma)$ by (1.4) with

$$h_j(k) = ka(j), \quad j \geq 1, \quad k \geq 0. \quad \dots (4.7)$$

By (4.3), we have $A(n) = o(\sqrt{n})$ and

$$B^2(n) = n \left(\int_{\mathbf{R}} x^2 d\Phi(x) + o(1) \right) = n(1 + o(1)).$$

By using (4.1) and (4.6), we conclude that the characteristic function $\varphi_n(t)$ of the distribution $\nu_n(h(\sigma) < x\sqrt{n})$ satisfies

$$\begin{aligned} \varphi_n(t) = & \frac{1}{2\pi i} \int_{1-Ki}^{1+iK} \frac{e^z}{z + t^2/2} dz \\ & + \left(1 + O\left(\frac{K}{n}\right) \right) \frac{1}{2\pi i} \int_{1-Ki}^{1+iK} \frac{e^z}{z + t^2/2} \left(\exp \left\{ \int_z^{z+t^2/2} \frac{e^{-w}}{w} dw \right\} - 1 \right) dz \\ & + O(K^{-1/2+\delta}). \end{aligned} \quad \dots (4.8)$$

The integrals occurring in (4.8) have been investigated in Timofeev (1985). The first of the integrals in (4.8) equals $e^{-t^2/2} + O(K^{-1})$. The bound $O((\log K)/K)$ for the second integral is obtained by changing the range of integration $[1 - iK, 1 + iK]$ using the contour $\{z : \Im z = \pm K, 1 \leq \Re z \leq K\}$ and $\{z : \Re z = K, |\Im z| \leq K\}$, and using the estimate

$$\exp \left\{ \int_z^{z+t^2/2} \frac{e^{-w}}{w} dw \right\} - 1 = O\left(\frac{e^{-\Re z}}{|z|}\right)$$

on the contour. Thus we obtain $\varphi_n(t) = e^{-t^2/2} + o(1)$ uniformly in $|t| \leq T$ for each $T \geq 1$. Consequently, we have

$$\nu_n(h(\sigma) < x\sqrt{n}) = \Phi(x) + o(1)$$

uniformly in $x \in \mathbf{R}$. This completes the proof of Theorem 2.

We conclude this paper with the conjecture: The process H_n described by h_j in (4.7) converges weakly in the space $\mathbf{D}[\mathbf{0}, \mathbf{1}]$, with the Skorokhod topology, to a process with dependent increments.

Appendix

Let $\Omega = \mathbf{Z}^{+n}$ denote the set of vectors with nonnegative components, $\omega = k_1\varepsilon_1 + \dots + k_n\varepsilon_n \in \Omega$ with $k_j = k_j(\omega) \geq 0$ and $\varepsilon_j = (0, \dots, 1, \dots, 0)$, where the j -th coordinate is 1, $1 \leq j \leq n$. We write $\omega \perp \eta$, if $k_1(\omega)k_1(\eta) + \dots + k_n(\omega)k_n(\eta) =$

0 and $\eta \leq \omega$ if $k_1(\eta) \leq k_1(\omega), \dots, k_n(\eta) \leq k_n(\omega)$. Let us adopt the notation $\eta \parallel \omega$ for the expression “ η exactly enters ω ” which means that $\eta \leq \omega$ and $\eta \perp \omega - \eta$. Let Ω be endowed with the product probability

$$P(\{\omega\}) = \prod_{j=1}^n p_j(k_j), \quad \sum_{k=0}^{\infty} p_j(k) = 1,$$

where $0 \leq p_j(k) \leq 1$ for $1 \leq j \leq n, k \geq 0$.

Let $(G, +)$ be an abelian additive group. A function $H : \Omega \rightarrow \mathbf{G}$ is called additive if it satisfies the condition $H(\alpha + \omega) = H(\alpha) + H(\omega)$ for each pair $\alpha, \omega \in \Omega$ and $\alpha \perp \omega$. Note that the additive function H has the representation

$$H(\omega) = \sum_{j=1}^n H(k_j \varepsilon_j), \quad H(0) = 0.$$

It shows that each double sequence $h_j(k) \in G$ with $h_j(0) = 0, 1 \leq j \leq n, k \geq 0$, defines via $H(k\varepsilon_j) = h_j(k)$ a unique additive function. The additive function satisfying, in addition, the condition $H(k_j \varepsilon_j) = k_j H(\varepsilon_j)$ for each $j \leq n$ is called a linear (completely additive) function.

Let $L : \Omega \rightarrow \mathbf{Z}^+$ be a linear function taking arbitrary positive values, except at $\omega = 0$. $\Omega_m = L^{-1}(m) = \{\omega : L(\omega) = m\}$. Lemma 1 motivates our interest in estimating of the probabilities

$$P(\{\omega : H(\omega) \in B\} | \Omega_n)$$

for $B \subset \mathbf{G}$ in terms of the unconditional ones.

For an arbitrary set $U \subset \Omega$ we define $\bar{U} = \Omega \setminus U$ and

$$V = V(U) := \{\omega = \alpha_1 + \alpha_2 - \alpha_3 : \alpha_1, \alpha_2, \alpha_3 \in U, \alpha_1 \perp \alpha_2 - \alpha_3, \alpha_3 \parallel \alpha_2\}.$$

LEMMA A. Suppose $n \geq 1$ and there exist positive constants c, c_1, C_1, C_2 such that

- (i) $p_j(0) \geq c$ for $1 \leq j \leq n$;
- (ii) $P(\Omega_m) \leq C_1 \left(\frac{n}{m+1}\right)^{1-\theta} P_n$ for $0 \leq m \leq n-1$ and for some $0 < \theta < 1$
- (iii) $P_n \geq c_1 n^{-1}$;
- (iv) for $1 \leq m \leq n$,

$$\sum_{\substack{k \geq 1, j \leq n \\ kL(\varepsilon_j) = m}} \frac{p_j(k)}{p_j(0)} \leq \frac{C_2}{m}.$$

Then

$$P(\bar{V} | \Omega_n) \leq CP^\theta(\bar{U}) + C_1 C_2 \theta^{-1} n^{-\theta},$$

where

$$C := \max \left\{ \frac{32}{c^2}, \frac{C_2}{c_1} + \frac{4C_1}{c} + \frac{C_1 C_2}{\theta} \right\}.$$

The result is nontrivial when $P_n := P(\Omega_n) = o(1)$. With some effort the estimate of C can be improved.

PROOF. Let Q_n be the set of the vectors $\pi := k\varepsilon_j$ with arbitrary $k \geq 1$ and $1 \leq j \leq n$ satisfying the condition $L(\pi) \leq n$ and put also $q(\pi) = p_j(k)/p_j(0)$ for such a $\pi \in Q_n$. Denote Q' the subset of Q_n consisting of π such that there exists an $\eta \in U$, $\eta \perp \pi$ and $\pi + \eta \in U$. Put $Q'' = Q_n \setminus Q'$. By Lemma 1 of Manstavičius (1998), we have

$$\sum_{\pi \in Q''} q(\pi) \leq 4c^{-1}P(\bar{U}) \tag{A.1}$$

provided $P(\bar{U}) < c^2/32$.

Further, let $\omega = \pi + \eta \in \bar{V} \cap \Omega_n$ with some $\pi = k\varepsilon_j$, $1 \leq L(\pi) \leq n$ such that $\pi \perp \eta$ and $L(\pi + \eta) = L(\pi) + L(\eta) = n$. Then $\pi \in Q''$ or $\eta \in \bar{U}$. Indeed, otherwise from the definition of Q' , there exists a vector $\eta_1 \in U$, $\eta_1 \perp \pi$ such that $\pi + \eta_1 \in U$ and $\omega = \eta + (\pi + \eta_1) - \eta_1 \in V$, which is impossible.

Since $P(\pi + \eta) = q(\pi)P(\eta)$, we obtain

$$\begin{aligned} P_n P(\bar{V} | \Omega_n) &= \frac{1}{n} \sum_{\omega \in \bar{V} \cap \Omega_n} P(\omega) \sum_{\pi \parallel \omega} L(\pi) \\ &= \frac{1}{n} \sum_{\substack{\pi + \eta \in \bar{V} \cap \Omega_n \\ \pi \perp \eta}} q(\pi)L(\pi)P(\eta) \\ &\leq \sum_{\pi \in Q''} q(\pi) \sum_{L(\eta)=n-L(\pi)} P(\eta) \\ &\quad + \frac{1}{n} \sum_{\eta \in \bar{U}} P(\eta)(n - L(\eta)) \sum_{L(\pi)=n-L(\eta)} q(\pi) \\ &=: \Sigma_1 + \Sigma_2. \end{aligned} \tag{A.2}$$

From conditions (iii) and (iv) we have

$$\Sigma_2 \leq \frac{C_2}{c_1} P_n P(\bar{U}).$$

We partition the sum Σ_1 into two sums. Let $0 < \delta < 1$ be an arbitrary parameter, then by (A.1) and condition (ii), we obtain

$$\sum_{\substack{\pi \in Q'' \\ L(\pi) \leq (1-\delta)n}} q(\pi)P(\Omega_{n-L(\pi)}) \leq C_1 \delta^{\theta-1} P_n \sum_{\pi \in Q''} q(\pi) \leq \frac{4C_1}{c} \delta^{\theta-1} P_n P(\bar{U}). \tag{A.3}$$

The remaining sum in Σ_1 is

$$\begin{aligned}
&\leq \sum_{(1-\delta)n < m \leq n} P(\Omega_{n-m}) \sum_{L(\pi)=m} q(\pi) \\
&\leq C_1 C_2 P_n \sum_{(1-\delta)n < m \leq n} \left(\frac{n}{n-m+1} \right)^{1-\theta} \frac{1}{m} \quad \dots (A.4) \\
&\leq C_1 C_2 P_n \frac{n^{1-\theta}}{(1-\delta)^n} \sum_{1 \leq j \leq \delta n+1} \frac{1}{j^{1-\theta}} \\
&\leq C_1 C_2 \theta^{-1} P_n n^{-\theta} (1+n\delta)^\theta \theta^{-1}.
\end{aligned}$$

Inequalities (A.3) and (A.4) together with the choice $\delta = P(\bar{U})$ imply that

$$\Sigma_1 \leq P_n \left(4C_1 c^{-1} P^\theta(\bar{U}) + C_1 C_2 \theta^{-1} P^\theta(\bar{U}) + C_1 C_2 \theta^{-1} n^{-\theta} \right).$$

Proof of Lemma A is completed by substituting the bounds of Σ_1 and Σ_2 into (A.2).

References

- ANTONIAK, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian non-parametric problems. *Ann. Statist.* **2** 1152–1174.
- ARRATIA, R., BARBOUR, A. D. AND TAVARÉ, S. (1992). Poisson process approximations for the Ewens sampling formula. *Ann. Applied Probab.* **2** 519–535.
- ARRATIA, R. AND TAVARÉ, S. (1992). Limit theorems for combinatorial structures via discrete process approximations. *Random Structures and Algorithms* **3** 321–345.
- BABU, G. J. (1973). A note on the invariance principle for additive functions. *Sankhyā A* **35** 307–310.
- BILLINGSLEY, P. (1968). *Convergence of Probability Measures*. John Wiley and Sons, 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.
- ERDÖS, P. AND TURÁN, P. (1965). On some problems of a statistical group theory I. *Z. Wahrsch. Verw. Gebiete* **4** 175–186.
- 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 (Russian).
- HANSEN, J. C. (1990). A functional central limit theorem for the Ewens Sampling Formula. *J. Appl. Probab.* **27** 28–43.
- JOHNSON, N. S., KOTZ, S. AND BALAKRISHNAN, N. (1997). *Discrete Multivariate Distributions*, Wiley, New York.
- 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.

- KOTZ, S., READ, S. AND BANKS, D. L. (1998). *Encyclopedia of Statistical Science*, Vol. 2, Wiley, New York.
- KUBILIUS, J. (1964). *Probabilistic Methods in the Theory of Numbers*. Ameri. Math. Soc. Translations **11** Providence, RI.
- MANSTAVIČIUS, E. (1984). Arithmetic simulation of stochastic processes. *Lith. Math. J.* **24** 276–285.
- — — (1985). Additive functions and stochastic processes. *Lith. Math. J.* **25** 52–61.
- — — (1996). Additive and multiplicative functions on random permutations. *Lith. Math. J.* **36** 400–408.
- — — (1998). The law of iterated logarithm for random permutations. *Lith. Matem. Rink.* **38** 205–220 (Russian).
- TIMOFEEV, N. M. AND USMANOV, H. H. (1982). Arithmetic simulation of the Brownian motion. *Dokl. Acad. Sci. Tadzh. SSR* **25** 207–211 (Russian).
- TIMOFEEV, N. M. (1985). Stable limit laws for additive arithmetic functions. *Matem. Zametki* **37** 465–473 (Russian).

GUTTI JOGESH BABU
DEPARTMENT OF STATISTICS
319 THOMAS BUILDING
THE PENNSYLVANIA STATE UNIVERSITY
UNIVERSITY PARK, PA 16802, USA
e-mail: babu@stat.psu.edu

EUGENIJUS MANSTAVIČIUS
DEPARTMENT OF MATHEMATICS
VILNIUS UNIVERSITY
VILNIUS
LITHUANIA
e-mail: eugenijus.manstavicius@maf.vu.lt