*The Annals of Applied Probability*
1992, Vol. 2, No. 3, 519–535

# POISSON PROCESS APPROXIMATIONS FOR THE EWENS SAMPLING FORMULA

BY RICHARD ARRATIA,[1] A. D. BARBOUR[2] AND SIMON TAVARÉ[1]

*University of Southern California, Universität Zürich and University of Southern California*

The Ewens sampling formula is a family of measures on permutations, that arises in population genetics, Bayesian statistics and many other applications. This family is indexed by a parameter $\theta > 0$; the usual uniform measure is included as the special case $\theta = 1$. Under the Ewens sampling formula with parameter $\theta$, the process of cycle counts $(C_1(n), C_2(n), \ldots, C_n(n), 0, 0, \ldots)$ converges to a Poisson process $(Z_1, Z_2, \ldots)$ with independent coordinates and $\mathbb{E}Z_j = \theta/j$. Exploiting a particular coupling, we give simple explicit upper bounds for the Wasserstein and total variation distances between the laws of $(C_1(n), \ldots, C_b(n))$ and $(Z_1, \ldots, Z_b)$. This Poisson approximation can be used to give simple proofs of limit theorems with bounds for a wide variety of functionals of such random permutations.

**1. Introduction.** The Ewens sampling formula with parameter $\theta > 0$ may be thought of as the measure on $\mathcal{S}_n$, the permutations of $\{1, 2, \ldots, n\}$, whose density with respect to uniform measure is proportional to $\theta^k$, where $k$ is the number of cycles in the permutation. The probability of the set of permutations having cycle index $(a_1, a_2, \ldots, a_n) \in \mathbf{Z}_+^n$ (i.e., having $a_j$ cycles of length $j$, for $j = 1, \ldots, n$) is

$$P_n(a_1, \ldots, a_n) = \frac{n!}{\theta_{(n)}} \prod_{j=1}^{n} \left(\frac{\theta}{j}\right)^{a_j} \frac{1}{a_j!} 1\left\{\sum_{l=1}^{n} l a_l = n\right\}, \tag{1}$$

where $\mathbf{Z}_+ \equiv \{0, 1, \ldots\}$ and

$$x_{(n)} \equiv x(x+1)\cdots(x+n-1), \qquad x_{(0)} = 1.$$

This formula was derived by Ewens (1972) in the context of population genetics, where $a_j$ is the number of alleles represented by $j$ genes in a sample of $n$ genes, and by Antoniak (1974) in a Bayesian nonparametric statistics setting. The special case $\theta = 1$ corresponds to each permutation being equally likely.

We let $C_j(n)$ be the number of cycles of size $j$ in an $n$-permutation. Under the Ewens sampling formula for fixed $\theta$, the finite-dimensional distributions of $(C_1(n), C_2(n), \ldots)$ converge to those of a Poisson process on $\mathbb{N} \equiv \{1, 2, \ldots\}$, as the following result shows.

Received February 1991; revised August 1991.
[1]Supported in part by NSF Grants DMS-88-15106, DMS-88-03284 and DMS-90-05833, and NIH Grant GM 41746.
[2]Supported in part by Schweizerischer NF Project 21-25579.88.
*AMS* 1980 *subject classifications*. 60C05, 05A05, 05A16, 92D10.
*Key words and phrases.* Total variation, population genetics, permutations.

THEOREM 1. *For $j = 1, 2, \ldots$ let $C_j(n)$ denote the number of cycles of length $j$ in an $n$-permutation following distribution $P_n$, so that $C_1(n) + 2C_2(n) + \cdots = n$. The process of cycle counts converges in distribution to a Poisson process on $\mathbb{N}$ with intensity $\theta/j$. That is, as $n \to \infty$,*

$$(C_1(n), C_2(n), \ldots) \Rightarrow (Z_1, Z_2, \ldots)$$

*where the $Z_j$, $j = 1, 2, \ldots$, are independent Poisson-distributed random variables with*

$$\mathbb{E}(Z_j) = \frac{\theta}{j}.$$

PROOF. For integers $0 \le l \le m$, define

$$T_{lm} = (l+1)Z_{l+1} + (l+2)Z_{l+2} + \cdots + mZ_m, \tag{2}$$

with $T_{mm} \equiv 0$. Choose $b \in \{1, \ldots, n\}$ and nonnegative integers $a_1, \ldots, a_b$ satisfying $a_1 + 2a_2 + \cdots + ba_b = a \le n$. Let $\mathbf{C}_b(n) = (C_1(n), \ldots, C_b(n))$, $\mathbf{Z}_b = (Z_1, \ldots, Z_b)$ and $\mathbf{a} = (a_1, \ldots, a_b)$. It follows from (1) [see also the discussion after (36)] that

$$\begin{aligned} \mathbb{P}(\mathbf{C}_b(n) = \mathbf{a}) &= \mathbb{P}(\mathbf{Z}_b = \mathbf{a} | T_{0n} = n) \\ &= \mathbb{P}(\mathbf{Z}_b = \mathbf{a}) \mathbb{P}(T_{bn} = n - a) / \mathbb{P}(T_{0n} = n), \end{aligned} \tag{3}$$

so the result will follow if we can show that as $n \to \infty$,

$$\mathbb{P}(T_{bn} = n - a) / \mathbb{P}(T_{0n} = n) \to 1. \tag{4}$$

The probability generating function of $T_{bn}$ is

$$\mathbb{E}x^{T_{bn}} = \exp\left( -\theta \sum_{j=b+1}^{n} 1/j \right) \exp\left( \theta \sum_{j=b+1}^{n} x^j/j \right).$$

If we define $g_{n-a} \equiv \exp(\theta \sum_{j=b+1}^{n} 1/j) \mathbb{P}(T_{bn} = n - a)$ and $f(x) \equiv \exp(-\theta \sum_{j=1}^{b} x^j/j)$, then

$$\begin{aligned} g_{n-a} &= [x^{n-a}] \exp\left( \theta \sum_{j=b+1}^{n} x^j/j \right) \\ &= [x^{n-a}] \exp\left( \theta \sum_{j=b+1}^{\infty} x^j/j \right) \\ &= [x^{n-a}] \exp\left( \theta \sum_{j=1}^{\infty} x^j/j \right) \exp\left( -\theta \sum_{j=1}^{b} x^j/j \right) \\ &= [x^{n-a}] (1-x)^{-\theta} f(x) \\ &= \frac{1}{(n-a)!} \left( f(1)\theta_{(n-a)} - f'(1)(\theta - 1)_{(n-a)} + \cdots \right) \\ &= \frac{\theta_{(n-a)}}{(n-a)!} \exp\left( -\theta \sum_{j=1}^{b} 1/j \right) \left( 1 + \frac{b\theta(\theta-1)}{\theta + n - a - 1} + O(n^{-2}) \right), \end{aligned} \tag{5}$$

the last two equalities following from Darboux's method [cf. Wilf (1990), Chapter 5]. Since

$$\mathbb{P}(T_{0n} = n) = \frac{\theta_{(n)}}{n!}\exp\left(-\theta \sum_{j=1}^{n} 1/j\right), \tag{6}$$

we see that as $n \to \infty$,

$$\frac{\mathbb{P}(T_{bn} = n - a)}{\mathbb{P}(T_{0n} = n)} = \frac{n!\theta_{(n-a)}}{(n-a)!\theta_{(n)}}(1 + O(n^{-1})).$$

Thus (4) holds, and the proof is complete. □

REMARK. This result may also be established by the method of moments using the results of Watterson (1974); the case $\theta = 1$ is described in Arratia and Tavaré (1992a). The present proof is included because the asymptotic expansion (5) is used later. In the case $\theta = 1$, the theorem is due to Kolchin (1971); see also Goncharov (1944).

In this paper we provide explicit estimates on the distance between the distribution of $(C_1(n), C_2(n), \dots)$ and the law of $(Z_1, Z_2, \dots)$, the independent Poisson components of Theorem 1. Specifically, for $1 \le b \le n$ we will estimate the distance $d_b^W(n)$ between $\mathbf{C}_b(n) \equiv (C_1(n), \dots, C_b(n))$ and $\mathbf{Z}_b \equiv (Z_1, \dots, Z_b)$ in the Wasserstein $l_1$ metric, defined by

$$d_b^W(n) = \inf \sum_{j=1}^{b} \mathbb{E}|C_j(n) - Z_j|, \tag{7}$$

where the infimum in (7) is taken over all couplings of $\mathbf{C}_b(n)$ and $\mathbf{Z}_b$ on a common probability space. In Theorem 2 we prove that

$$d_b^W(n) \le \frac{b\theta}{\theta + n - b}\left(\theta + \frac{n}{\theta + n}\right), \tag{8}$$

and that $d_n^W(n)$ is uniformly bounded in $n$. The use of these estimates to prove limit theorems for functionals of the cycle counting process is the subject of Arratia and Tavaré (1992b).

The coupling exploited in the proof of Theorem 2 may also be used to estimate $d_b(n)$, the total variation distance between the law of $\mathbf{C}_b(n)$ and the law of $\mathbf{Z}_b$, defined by

$$\begin{aligned} d_b(n) &\equiv \|\mathcal{L}(\mathbf{C}_b(n)) - \mathcal{L}(\mathbf{Z}_b)\| \\ &= \sup_{A \subseteq \mathbf{Z}_+^b} |\mathbb{P}(\mathbf{C}_b(n) \in A) - \mathbb{P}(\mathbf{Z}_b \in A)| \qquad (9) \\ &= \inf_{\text{couplings}} \mathbb{P}(\mathbf{C}_b(n) \ne \mathbf{Z}_b). \qquad (10) \end{aligned}$$

In Theorem 3 we prove that $d_b(n) \to 0$ if and only if $b = o(n)$, and that in any case,

$$d_b(n) \le \frac{b\theta}{\theta + n}\left(\theta + \frac{n}{\theta + n - b}\right). \tag{11}$$

For the case $\theta = 1$, the result $d_b(n) \leq 2b/n$ was proved by Diaconis and Pitman (1986) and independently by Barbour (1990). When $\theta = 1$, the bound in (11) may be much improved. Arratia and Tavaré (1992a) show that if $b/n \to 0$, then $d_b(n) \to 0$ superexponentially fast relative to $n/b$.

**2. The Chinese restaurant process.** There are several couplings of permutations of $\{1, 2, \ldots, n\}$ for different values of $n$ that preserve interesting features of the cycle structure. One of these is known as the "Chinese restaurant process" [Dubins and Pitman, quoted in Aldous (1985)]; a second is the Feller coupling.

The Chinese restaurant process generates permutations of size $1, 2, \ldots$ sequentially, as follows. Define independent random variables $A_1, A_2, \ldots$ with distributions determined by

$$\mathbb{P}(A_i = j) = \begin{cases} \dfrac{\theta}{\theta + i - 1}, & j = i, \\ \dfrac{1}{\theta + i - 1}, & j = 1, 2, \ldots, i - 1. \end{cases} \tag{12}$$

The sequence $A_1, A_2, \ldots$ is used sequentially to generate the cycles of a permutation. The integer 1 starts a cycle. The integer 2 is placed to the right of 1 (in the same cycle) with probability $1/(\theta + 1)$, or begins a new cycle with probability $\theta/(\theta + 1)$. Suppose then that the first $n - 1$ integers have been assigned to cycles. Integer $n$ starts a new cycle with probability $\mathbb{P}(A_n = n) = \theta/(\theta + n - 1)$, or is placed to the right of integer $j$ with probability $\mathbb{P}(A_n = j) = 1/(\theta + n - 1)$, $j = 1, \ldots, n - 1$. A simple calculation shows that for any $\pi \in \mathcal{S}_n$ having $k$ cycles,

$$\mathbb{P}(\pi) = \frac{\theta^k}{\theta_{(n)}},$$

so that the probability of the set of permutations with cycle index $(a_1, \ldots, a_n)$ is given by (1); see Joyce and Tavaré (1987) and Diaconis and Pitman (1986). Note that there is an ordering of the cycles produced this way; the first cycle contains the integer 1, the second cycle contains the smallest integer not in the first cycle and so on.

The Chinese restaurant process is an elaboration of couplings which generate only the partition associated with the cycle decomposition of a permutation. These couplings for the partition structure appeared in several guises. Bollobás (1985), Theorem 25, discusses the case of uniform random permutations. Blackwell and McQueen (1973) and Hoppe (1984) address the connection with certain urn models. Ewens (1972), Hoppe (1984), Donnelly (1986), Donnelly and Tavaré (1986), Hoppe (1987) and Tavaré (1987) discuss applications to population genetics, where several approaches to the study of large cycles (i.e., those with size comparable to $n$) are described, albeit in the language of population genetics. Some other couplings are discussed in Goldie (1989).

This coupling has the property that once the integers $i$ and $j$ are assigned to a common cycle, they remain so for ever after. However, the cycle lengths themselves change continually. In the next section we describe an alternative coupling in which cycle lengths remain frozen as $n$ increases.

**3. The Feller coupling.** Let $\xi_1, \xi_2, \ldots$ be independent Bernoulli random variables with distribution given by

$$p_j \equiv \mathbb{P}(\xi_j = 1) = \frac{\theta}{\theta + j - 1}, \qquad j = 1, 2, \ldots . \tag{13}$$

We use the $\xi$ sequence in the order $\xi_n, \xi_{n-1}, \ldots, \xi_1$ to construct an $n$ permutation with ordered cycles as follows. Start with 1 in the first cycle. If $\xi_n = 1$, we finish that cycle and start the next with the smallest available integer. If, on the other hand, $\xi_n = 0$ we choose one of the remaining $n - 1$ integers at random and place it to the right of 1 in the same cycle. Continuing in this way produces an $n$ permutation with cycles ordered by their smallest integer. In terms of the independent auxiliary random variables $A_i$ defined in (12), we may define $\xi_i = 1(A_i = i)$, and on the event $\{A_i < i\}$, $A_i$ specifies which of the remaining $i - 1$ integers is used next. Note that to construct an $n$-permutation, the Chinese restaurant process uses $A_1, A_2, \ldots, A_n$, while the Feller coupling uses the reverse order, $A_n, \ldots, A_1$. For any $\pi \in \mathcal{S}_n$ with $k$ cycles, it is immediate that $\mathbb{P}(\pi) = \theta^k/\theta_{(n)}$, so that once more the probability that a permutation has cycle index $(a_1, \ldots, a_n)$ is given by (1). In this construction, which elaborates on an idea of Feller (1945), the cycles are built and completed one by one, in contrast to the Chinese restaurant process. Since the number of cycles produced by this construction is precisely $\sum_{j=1}^{n} \xi_j$, it follows that

$$\sum_{j=1}^{n} C_j(n) = \sum_{j=1}^{n} \xi_j.$$

We can think of the lengths of the first cycle, the second cycle, ... as the spacings between consecutive ones in the sequence $1, \xi_n, \ldots, \xi_1$. It follows that the number $C_j(n)$ of cycles of length $j$ is

$$\begin{aligned} C_j(n) = &\sum_{i=1}^{n-j} \xi_i(1 - \xi_{i+1}) \cdots (1 - \xi_{i+j-1})\xi_{i+j} \\ &+ \xi_{n-j+1}(1 - \xi_{n-j+2}) \cdots (1 - \xi_n). \end{aligned} \tag{14}$$

Since $(C_1(n), \ldots, C_n(n))$ has distribution (1), it follows from Theorem 1 that $(C_1(n), C_2(n), \ldots) \Rightarrow (Z_1, Z_2, \ldots)$ as $n \to \infty$. Next define $C_j(\infty)$ to be the number of $j$-spacings in the sequence $\xi_1, \xi_2, \ldots$ ; that is,

$$C_j(\infty) = \sum_{i=1}^{\infty} \xi_i(1 - \xi_{i+1}) \cdots (1 - \xi_{i+j-1})\xi_{i+j}. \tag{15}$$

Since $\mathbb{E}C_j(\infty) = \theta/j$, each $C_j(\infty)$ is almost surely finite. In fact, their joint distribution is easy to find. Since

$$\mathbb{E}\xi_{n-j+1}(1-\xi_{n-j+2})\cdots(1-\xi_n) \le \theta/(\theta+n-j+1) \to 0 \quad \text{as} \quad n \to \infty$$

for fixed $j$, it follows from (14) that $(C_1(n), C_2(n), \dots) \Rightarrow (C_1(\infty), C_2(\infty), \dots) \stackrel{d}{=} (Z_1, Z_2, \dots)$. Thus the $C_j(\infty)$ are independent Poisson random variables with $\mathbb{E}C_j(\infty) = \theta/j$.

We have now constructed a coupling of $(C_1(n), C_2(n), \dots)$ and $(Z_1, Z_2, \dots)$ for all $n$ simultaneously. Before estimating how close the coupling is, we highlight one aspect that is very useful in applications; for example, Arratia and Tavaré (1992b) use Lemma 1 to provide a simple proof of the Erdös–Turán law [Erdös and Turán (1967)]. The proof of Lemma 1 is an immediate consequence of the definitions in (14) and (15) and is omitted.

LEMMA 1. *For each $n$ and $1 \le j \le n$, we have*

$$C_j(n) \le C_j(\infty) + 1(J_n = j), \tag{16}$$

*where the random variable $J_n \in \{1, 2, \dots, n\}$ is defined by*

$$J_n = \min\{j \ge 1 : \xi_{n+1-j} = 1\}. \tag{17}$$

REMARK. Properties of distributions involving $J_n$ are often readily computed using (17) and the independent Bernoulli structure of the $\{\xi_j\}$. In particular,

$$\begin{aligned}\mathbb{P}(J_n = j) &= \mathbb{P}(\xi_{n-j+1} = 1, \xi_{n-j+2} = \cdots = \xi_n = 0)\\ &= \frac{\theta(n-1)!\Gamma(n-j+\theta)}{(n-j)!\Gamma(n+\theta)}\\ &= \frac{j\mathbb{E}C_j(n)}{n}.\end{aligned}$$

The random variable $J_n$ is familiar in the population genetics context, where it represents the number of alleles of the oldest allelic type in a sample of $n$ genes. See Kelly (1977) and Donnelly and Tavaré (1986), for example.

The Feller coupling provides a natural setting for the study of large cycles. We do not pursue this aspect here [see Ignatov (1981) and Arratia, Barbour and Tavaré (1992)], but instead concentrate on the behavior of the small cycles.

**4. Upper bounds via the Feller coupling.** One use of couplings is to obtain upper bounds on distances between probability measures. We begin by giving upper bounds on $d_b^W(n)$, and showing that $d_n^W(n)$ is uniformly bounded in $n$.

THEOREM 2. *For all* $1 \le b \le n$,

$$d_b^W(n) \le \frac{b\theta}{\theta + n - b}\left(\theta + \frac{n}{\theta + n}\right) \le \frac{b\theta(\theta + 1)}{\theta + n - b}, \tag{18}$$

*and if* $\theta \ge 1$, *we also have*

$$d_b^W(n) \le \frac{b\theta(\theta + 1)}{n + \theta}. \tag{19}$$

*Furthermore, for any* $\theta > 0$, $\{d_n^W(n), n = 1, 2, \ldots\}$ *is a bounded sequence.*

PROOF. To establish the inequality (18), we use the fact that

$$\begin{aligned} C_j(\infty) - C_j(n) = &\sum_{l > n-j+1} \xi_l(1 - \xi_{l+1}) \cdots (1 - \xi_{l+j-1})\xi_{l+j} \\ &- \xi_{n-j+1}(1 - \xi_{n-j+2}) \cdots (1 - \xi_{n+1}). \end{aligned} \tag{20}$$

Leaving out the intermediate $j - 1$ factors from each term,

$$\begin{aligned} \mathbb{E}\left|C_j(\infty) - C_j(n)\right| &\le \sum_{l > n-j+1} p_l p_{l+j} + p_{n-j+1}(1 - p_{n+1}) \\ &= \theta^2 \sum_{l > n-j+1} \frac{1}{(\theta + l - 1)(\theta + l + j - 1)} \\ &\quad + \frac{\theta}{(\theta + n - j)} \frac{n}{(n + \theta)} \\ &= \frac{\theta^2}{j} \sum_{l=n-j+1}^{n} \frac{1}{\theta + l} + \frac{\theta}{(\theta + n - j)} \frac{n}{(n + \theta)} \\ &\le \frac{\theta^2}{\theta + n - j + 1} + \frac{\theta}{(\theta + n - j)} \frac{n}{(n + \theta)}. \end{aligned}$$

From this last inequality, we see that

$$\begin{aligned} \sum_{j=1}^{b} \mathbb{E}\left|C_j(\infty) - C_j(n)\right| &\le \frac{\theta^2 b}{\theta + n - b + 1} + \frac{n\theta}{(\theta + n)} \frac{b}{(\theta + n - b)} \\ &\le \frac{\theta b}{\theta + n - b}\left(\theta + \frac{n}{\theta + n}\right). \end{aligned}$$

When $\theta \geq 1$, we may use (20) and (13) to see that

$$\begin{aligned}
\mathbb{E}\left|C_j(\infty) - C_j(n)\right| &\leq \sum_{l>n-j+1} \frac{\theta}{(\theta+l-1)} \frac{l}{(l+\theta)} \\
&\quad \cdots \frac{l+j-2}{(\theta+l+j-2)} \frac{\theta}{(\theta+l+j-1)} \\
&\quad + \frac{\theta}{(n-j+\theta)} \frac{n-j+1}{(\theta+n-j+1)} \cdots \frac{n}{(\theta+n)} \qquad (21) \\
&\leq \theta^2 \sum_{l>n-j+1} \frac{1}{(\theta+l+j-2)(\theta+l+j-1)} + \frac{\theta}{\theta+n} \\
&= \frac{\theta^2}{\theta+n} + \frac{\theta}{\theta+n}.
\end{aligned}$$

Summing the inequality over $j = 1, \ldots, b$ shows that

$$d_b^W(n) \leq \frac{b\theta(\theta+1)}{\theta+n}.$$

To show that the sequence $d_n^W(n)$ is uniformly bounded, observe that for any $1 \leq b \leq n$,

$$\begin{aligned}
d_n^W(n) &\leq \sum_{j=1}^{n} \mathbb{E}\left|C_j(\infty) - C_j(n)\right| \\
&\leq \frac{b\theta(\theta+1)}{\theta+n-b} + \sum_{j=b+1}^{n} \mathbb{E}C_j(\infty) + \sum_{j=b+1}^{n} \mathbb{E}C_j(n).
\end{aligned}$$

From (16), the rightmost two terms are at most $1 + 2\sum_{j=b+1}^{n} \mathbb{E}C_j(\infty)$, which in turn is at most $1 + 2\log(n/b)$. Choosing $b = \lfloor n/2 \rfloor$, it follows that for $n \geq 3$,

$$d_n^W(n) \leq 1 + \frac{n\theta(\theta+1)}{2\theta+n} + 2\log\left(\frac{2n}{n-2}\right) \to 1 + \theta(\theta+1) + 2\log 2,$$

which completes the proof of the theorem. □

The next result provides an upper bound on $d_b(n)$ that is also $O(b/n)$.

THEOREM 3. *As $n \to \infty$, $d_b(n) \to 0$ if and only if $b/n \to 0$. For all $1 \leq b \leq n$,*

$$d_b(n) \leq \frac{b\theta}{\theta+n}\left(\theta + \frac{n}{\theta+n-b}\right). \qquad (22)$$

*Furthermore, if $\theta \geq 1$ and $1 \leq b \leq n$,*

$$d_b(n) \leq \frac{b\theta(\theta+1)}{\theta+n}. \qquad (23)$$

PROOF. The following proof of the "only if" statement is analogous to the proof of the corresponding statement in Theorem 2 of Arratia and Tavaré (1992a) for the case $\theta = 1$. We will show that if $b/n \geq \varepsilon > 0$ for all $n, b$ then $\liminf_{n\to\infty} d_b(n) > 0$. From (10), note that

$$\begin{aligned} d_b(n) &\geq \mathbb{P}(T_{0b} > n) \\ &\geq \mathbb{P}\left(\sum_{b/2<i\leq b} iZ_i > n\right) \\ &\geq \mathbb{P}\left(\frac{b}{2}\sum_{b/2<i\leq b} Z_i > n\right) \\ &\geq \mathbb{P}\left(\sum_{b/2<i\leq b} Z_i > \frac{2}{\varepsilon}\right) \\ &\to \mathbb{P}\left(\text{Poisson}(\theta \log 2) > \frac{2}{\varepsilon}\right) \\ &> 0. \end{aligned}$$

If $b/n \geq \varepsilon > 0$ for infinitely many $n$, we may apply the argument above to an appropriate subsequence to establish the "only if" part of the theorem.

To establish the bound (22), consider the event

$$E = \{(C_1(n), \ldots, C_b(n)) = (C_1(\infty), \ldots, C_b(\infty))\},$$

and recall from (10) that

$$d_b(n) \leq \mathbb{P}(E^c).$$

To estimate $\mathbb{P}(E^c)$, observe that

$$\begin{aligned} E \supseteq &(\{\xi_{n+1} = 1\} \cup \{\xi_{n-b+1} = \cdots = \xi_n = 0\}) \\ &\cap \bigcap_{j>n} (\{\xi_j = 0\} \cup \{\xi_{j+1} = \cdots = \xi_{j+b} = 0\}). \end{aligned}$$

Now use (13) to see that for any $l \geq 1$,

$$\begin{aligned} \mathbb{P}(\{\xi_{l+1} = \cdots = \xi_{l+b} = 0\}^c) &= \mathbb{P}\left(\bigcup_{m=1}^{b} \{\xi_{l+m} = 1\}\right) \\ &\leq \sum_{m=1}^{b} \mathbb{P}(\xi_{l+m} = 1) \\ &\leq \frac{b\theta}{\theta + l}. \end{aligned} \tag{24}$$

Hence

$$\begin{aligned}\mathbb{P}(E^c) &\le \frac{n}{(n+\theta)}\frac{b\theta}{(\theta+n-b)} + \sum_{j>n}\frac{\theta}{(\theta+j-1)}\frac{b\theta}{(\theta+j)} \\ &= \frac{n}{(n+\theta)}\frac{b\theta}{(\theta+n-b)} + \frac{b\theta^2}{(\theta+n)},\end{aligned} \tag{25}$$

which establishes the bound in (22).

To establish (23), note from (19) that when $\theta \ge 1$,

$$d_b(n) \le d_b^W(n) \le \frac{b\theta(\theta+1)}{\theta+n}. \qquad \square$$

REMARKS. Comparing the bounds on $d_b$ in (22) and (23), we find that the bound in (23) is sharper for $b > \theta$, and in particular as $b \to \infty$. When $\theta = 1$, $d_b(n) \le 2b/(n+1)$, recovering the earlier result of Diaconis and Pitman (1986) and Barbour (1990).

**5. Lower bounds.** Lower bounds for distances between probability measures cannot be deduced from a particular coupling, but require special arguments, such as the following. We begin by establishing that if $\theta \ne 1$, then $nd_b^W(n)/b$ is bounded away from 0.

The definition of $d_b^W(n)$ in (7),

$$d_b^W(n) = \inf_{\text{couplings}} \sum_{j=1}^{b} \mathbb{E}\left|C_j(n) - Z_j\right|,$$

shows that

$$d_b^W(n) \ge \sum_{j=1}^{b} \left|\mathbb{E}C_j(n) - \mathbb{E}Z_j\right|, \tag{26}$$

a result that may be used to establish the following lower bounds on $d_b^W(n)$, valid for all $1 \le b \le n$.

THEOREM 4. *For all $b \le n$, and for $\theta \ge 1$, we have*

$$d_b^W(n) \ge \frac{\theta(\theta-1)b}{\theta+n-1} - \frac{\theta(\theta-1)^2 b(b+1)}{4(\theta+n-1)^2}, \tag{27}$$

*while if $\theta \le 1$,*

$$d_b^W(n) \ge \frac{\theta(1-\theta)b}{\theta+n-1}. \tag{28}$$

PROOF. Watterson (1974) established that

$$\mathbb{E}C_j(n) = \frac{\theta}{j}\frac{n(n-1)\cdots(n-j+1)}{(\theta+n-j)\cdots(\theta+n-1)}. \tag{29}$$

In the case that $\theta \geq 1$, we have

$$\begin{aligned}
\mathbb{E}Z_j - \mathbb{E}C_j(n) &= \frac{\theta}{j}\left(1 - \prod_{l=1}^{j}\left(1 - \frac{\theta-1}{\theta+n-l}\right)\right) \\
&\geq \frac{\theta}{j}\left(1 - \exp\left(-(\theta-1)\sum_{l=1}^{j}\frac{1}{\theta+n-l}\right)\right) \\
&\geq \frac{\theta}{j}\left(1 - \exp\left(-\frac{(\theta-1)j}{(\theta+n-1)}\right)\right) \qquad (30) \\
&\geq \frac{\theta}{j}\left(\frac{(\theta-1)j}{(\theta+n-1)} - \frac{(\theta-1)^2 j^2}{2(\theta+n-1)^2}\right) \\
&= \frac{\theta(\theta-1)}{\theta+n-1} - \frac{\theta(\theta-1)^2 j}{2(\theta+n-1)^2}.
\end{aligned}$$

Summing the last inequality over $j = 1, \ldots, b$ completes the proof of (27).

On the other hand, if $\theta \leq 1$ we have

$$\begin{aligned}
\mathbb{E}C_j(n) - \mathbb{E}Z_j &= \frac{\theta}{j}\left(\prod_{l=1}^{j}\left(1 + \frac{1-\theta}{\theta+n-1}\right) - 1\right) \\
&\geq \frac{\theta}{j}\sum_{l=1}^{j}\frac{1-\theta}{\theta+n-l} \\
&\geq \frac{\theta}{j}\frac{(1-\theta)j}{\theta+n-1} \\
&= \frac{\theta(1-\theta)}{\theta+n-1}.
\end{aligned}$$

Summing this inequality over $j = 1, \ldots, b$ establishes (28), and completes the proof. □

Establishing so sharp a lower bound on $d_b(n)$ is rather more difficult. We will derive lower bounds of order $O(((n/b)\log(n/b))^{-1})$ valid as $b, n \to \infty$. The method is based on the following result.

LEMMA 2. *Let $X$ and $Y$ be nonnegative random variables such that $\mathbb{E}X - \mathbb{E}Y = \varepsilon > 0$ and $\mathbb{E}X^r \equiv m_r < \infty$ for some $r > 1$. Then*

$$\|\mathscr{L}(X) - \mathscr{L}(Y)\| \geq \left[\left(\frac{r-1}{r}\right)\left(\frac{\varepsilon}{m_r^{1/r}}\right)\right]^{r/(r-1)} \tag{31}$$

PROOF. Clearly,

$$\inf_{Z:\, \mathbb{E}Z = \mathbb{E}X - \varepsilon} \|\mathscr{L}(X) - \mathscr{L}(Z)\|$$

is attained by $Z = X^\varepsilon$, where $X^\varepsilon = X1\{X \leq x(\varepsilon)\}$, with $x(\varepsilon)$ chosen so that $\mathbb{E}X^\varepsilon = \mathbb{E}X - \varepsilon$ [randomizing if necessary on the set $\{X = x(\varepsilon)\}$]. Furthermore, if $W$ is any random variable such that $W \geq X$, and $W^\varepsilon$ is the corresponding truncation of $W$, with $\mathbb{E}W^\varepsilon = \mathbb{E}W - \varepsilon$, it is immediate that

$$\|\mathscr{L}(W) - \mathscr{L}(W^\varepsilon)\| \leq \|\mathscr{L}(X) - \mathscr{L}(X^\varepsilon)\|.$$

So choose a random variable $W \geq X$ with distribution given by

$$\mathbb{P}(W \geq w) = \min\{1, m_r w^{-r}\}, \qquad w \geq 0,$$

which can be done because of Markov's inequality. The truncation level $w(\varepsilon)$ for $W$ satisfies

$$\varepsilon = EW - \mathbb{E}W^\varepsilon = \int_{w(\varepsilon)}^{\infty} r m_r w^{-(r+1)} w \, dw = \{r m_r/(r-1)\}[w(\varepsilon)]^{-r+1}.$$

Observe that

$$\mathbb{P}(W^\varepsilon \neq W) = \mathbb{P}(W \geq w(\varepsilon)) = m_r[w(\varepsilon)]^{-r}.$$

This implies that

$$\begin{aligned}\|\mathscr{L}(X) - \mathscr{L}(Y)\| &\geq \|\mathscr{L}(X) - \mathscr{L}(X^\varepsilon)\| \\ &\geq \|\mathscr{L}(W) - \mathscr{L}(W^\varepsilon)\| \\ &= m_r\left(\frac{(r-1)\varepsilon}{r m_r}\right)^{r/(r-1)},\end{aligned}$$

proving the lemma. □

COROLLARY 1. *There exists a constant $c > 0$ such that if $X \sim$ Poisson$(\lambda)$ and $Y$ is any nonnegative random variable with $\mathbb{E}Y = \lambda - \varepsilon$ and $0 < \varepsilon < e^{-2}$, then*

$$\|\mathscr{L}(X) - \mathscr{L}(Y)\| \geq \frac{c\varepsilon}{\max\{\lambda e, \log(1/\varepsilon)\}}.$$

PROOF. Observe that if $X \sim$ Poisson$(\lambda)$, then

$$m_r \leq r^r + (\lambda e)^r. \tag{32}$$

Apply the previous theorem with $r = \max\{\lambda e, \log(1/\varepsilon)\}$, and note that

$$\left(\frac{\varepsilon}{r}\right)^{1/(r-1)} > \left(\frac{e^{-\lambda e}}{\lambda e}\right)^{1/(\lambda e-1)} \wedge \left(\frac{\varepsilon}{-\log \varepsilon}\right)^{1/(-\log \varepsilon - 1)},$$

which is bounded away from 0 for all $0 < \varepsilon < e^{-2}$ and all $\lambda > 2/e$. □

We will use Lemma 2 and Corollary 1 to establish the following result.

THEOREM 5. *If $\theta \neq 1$ and $b, n \to \infty$, then $(n/b)\log(n/b)\, d_b(n)$ is bounded away from 0.*

PROOF. Observe first that for any $f\colon \mathbb{N}^b \to \mathbb{R}^r$ with $r < b$,

$$\|\mathscr{L}(\mathbf{C}_b(n)) - \mathscr{L}(\mathbf{Z}_b)\| \geq \|\mathscr{L}f(\mathbf{C}_b(n)) - \mathscr{L}f(\mathbf{Z}_b)\|.$$

In particular, for any $L \leq b$,

$$\|\mathscr{L}(\mathbf{C}_b(n)) - \mathscr{L}(\mathbf{Z}_b)\| \geq \|\mathscr{L}(C_L(n) + \cdots + C_b(n)) - \mathscr{L}(Z_L + \cdots + Z_b)\|.$$

First, we treat the case $\theta > 1$. We will choose $L = \lfloor b/e \rfloor$, and use Corollary 1 with $X = Z_L + \cdots + Z_b$, $Y = C_L(n) + \cdots + C_b(n)$. In this case,

$$\lambda = \mathbb{E}X = \sum_{j=L}^{b} \frac{\theta}{j} \sim \theta, \qquad b \to \infty. \tag{33}$$

We may assume that $b/n \to 0$ as $n \to \infty$. A lower bound on $\varepsilon \equiv \sum_{j=L}^{b}(\mathbb{E}Z_j - \mathbb{E}C_j(n))$ may be found from (30), from which it follows that there exists a constant $c_1 > 0$ such that $\varepsilon > c_1 b/n$ for sufficiently large $n$. To establish an upper bound on $\varepsilon$, note that from (30),

$$\begin{aligned}
\mathbb{E}Z_j - \mathbb{E}C_j(n) &= \frac{\theta}{j}\left(1 - \prod_{l=1}^{j}\left(1 - \frac{\theta - 1}{\theta + n - l}\right)\right) \\
&\leq \frac{\theta}{j}\sum_{l=1}^{j}\frac{\theta - 1}{\theta + n - l} \\
&\leq \frac{\theta(\theta - 1)}{\theta + n - j}.
\end{aligned}$$

It follows that there exists a constant $c_2$ such that $\varepsilon \leq c_2 b/n$ for sufficiently large $n$. In summary, there are constants $0 < c_1 < c_2$ such that $c_1 b/n \leq \varepsilon \leq c_2 b/n$ for sufficiently large $n$. Since $\lambda \sim \theta$ and $\varepsilon \to 0$, it follows from Corollary 1 that there is a constant $c$ such that

$$d_b(n) \geq c\varepsilon/(-\log \varepsilon),$$

from which the result follows.

In the case $0 < \theta < 1$, a different argument is required. If $\mathcal{S}_n \subseteq \{1, \ldots, n\}$, the inequality (16) shows that

$$\sum_{j \in \mathcal{S}_n} C_j(n) \leq 1 + \sum_{j \in \mathcal{S}_n} C_j(\infty). \tag{34}$$

We will choose $L = \lfloor b/2 \rfloor + 1$, and use Theorem 2 with $X = C_L(n) + \cdots + C_b(n)$ and $Y = Z_L + \cdots + Z_b$. Note that $Y$ has a Poisson distribution with mean $\mathbb{E}Y = \theta \sum_{j=L}^{b} 1/j \leq \theta \log 2$. Since $\{C_j(\infty),\ j \geq 1\}$ has the same law as $\{Z_j,\ j \geq 1\}$, it follows from (34) that

$$\begin{aligned} m_r &\equiv \mathbb{E}X^r \\ &\leq \mathbb{E}(Y+1)^r \\ &\leq (1+r)^r + (1 + e\mathbb{E}Y)^r \\ &\leq (1+r)^r + (1 + e\log 2)^r, \end{aligned} \tag{35}$$

the last but one inequality following from (32). As in the earlier part of the proof, there exist constants $0 < c_1 < c_2$ such that $\varepsilon \equiv \sum_{j=L}^{b} (\mathbb{E}C_j(n) - \mathbb{E}Z_j)$ satisfies $c_1 b/n \leq \varepsilon \leq c_2 b/n$ for sufficiently large $n$ (with $b/n \to 0$). Since $\varepsilon \to 0$, we may take $r = \log(1/\varepsilon)$ in Lemma 2 to see that there is a constant $c$ such that $d_b(n) \geq c\varepsilon/(-\log \varepsilon)$, completing the proof of the theorem. □

**6. Discussion.** The total variation distance between the process $(C_1(n), \ldots, C_b(n))$ and the process $(Z_1, \ldots, Z_b)$ can be expressed in terms of the total variation distance between two random variables as follows:

$$\begin{aligned} d_b(n) &= \| \mathcal{L}(T_{0b}) - \mathcal{L}(T_{0b} | T_{0n} = n) \| \\ &= \frac{1}{2} \sum_{r=0}^{\infty} \mathbb{P}(T_{0b} = r) \left| \frac{\mathbb{P}(T_{bn} = n - r)}{\mathbb{P}(T_{0n} = n)} - 1 \right|. \end{aligned} \tag{36}$$

Relation (36) is given as Lemmas 1 and 9 in Arratia and Tavaré (1992a) for uniform measure on permutations and on other combinatorial assemblies. The extension to the Ewens sampling formula is straightforward. For the case $\theta = 1$, Arratia and Tavaré (1992a) analyze the above relation to show that $d_b(n)$ decays superexponentially fast relative to $n/b \to \infty$. Theorem 5 showed that if $\theta \neq 1$, we cannot hope for superexponential decay for $d_b(n)$ for fixed $b$ as $n \to \infty$. The intuitive reason that superexponential decay only occurs for $\theta = 1$ is that (36) involves conditioning on the event $T_{0n} = n$, where $\mathbb{E}T_{0n} = \theta_n$; conditioning on the event that something equals its mean, which is the case $\theta = 1$, is qualitatively different from conditioning on something being a fixed nonzero number of standard deviations away from the mean, corresponding to any case with $\theta \neq 1$.

In Theorem 6, we provide a lower bound for the total variation distance $d_b(n) \equiv \| \mathcal{L}(\mathbf{C}_b(n)) - \mathcal{L}(\mathbf{Z}_b) \|$ that is $O(b/n)$ for fixed $b$ as $n \to \infty$.

THEOREM 6. *For fixed* $b$,

$$\liminf_{n\to\infty} nd_b(n) \geq \frac{b\theta|\theta - 1|}{2} \exp\left(-\theta \sum_{j=1}^{b} 1/j\right).$$

*In particular, if* $\theta \neq 1$, *then* $\liminf_{n\to\infty} nd_b(n) > 0$.

PROOF. From (36) we see that

$$d_b(n) \geq \frac{1}{2}\mathbb{P}(T_{0b} = 0)\left|\frac{\mathbb{P}(T_{bn} = n)}{\mathbb{P}(T_{0n} = n)} - 1\right|.$$

We will use Darboux's method once more to estimate $\mathbb{P}(T_{bn} = n)$. From (5) with $a = 0$, we see that

$$\mathbb{P}(T_{bn} = n) = \exp\left(-\theta \sum_{j=1}^{n} 1/j\right) \frac{\theta_{(n)}}{n!}\left(1 + \frac{b\theta(\theta - 1)}{\theta + n - 1} + O(n^{-2})\right),$$

so that from (6),

$$\frac{\mathbb{P}(T_{bn} = n)}{\mathbb{P}(T_{0n} = n)} - 1 = \frac{b\theta(\theta - 1)}{\theta + n - 1} + O(n^{-2}).$$

The proof is now completed by recalling that $\mathbb{P}(T_{0b} = 0) = \exp(-\theta\Sigma_{j=1}^{b} 1/j)$. □

REMARK. One would expect on the basis of Theorem 6 that the correct order for the lower bound on $d_b(n)$ is $O(b/n)$. This has now been proved in Barbour (1992).

We conclude with a brief discussion of the asymptotic behavior of $d_n(n)$. Equations (36) and (6) show that

$$\begin{aligned} d_n(n) &= \mathbb{P}(T_{0n} \neq n) \\ &= 1 - \frac{\theta_{(n)}}{n!} \exp\left(-\sum_{j=1}^{n} \frac{\theta}{j}\right), \end{aligned}$$

so that as $n \to \infty$,

$$d_n(n) = 1 - \frac{e^{-\theta\gamma}}{\Gamma(\theta)n}(1 + o(1)),$$

where $\gamma$ is Euler's constant. For other values of $n$ and $b$, we have computed $d_b(n)$ numerically using (36) and the REDUCE computer algebra program [Hearn (1987); see Arratia and Tavaré (1992a) for details]. Table 1 gives the values of $d_b(n)$ for $n = 10, 50, 100, 250$ and $b = 1, \lfloor n^{1/3} \rfloor, \lfloor n^{1/2} \rfloor$.

Watterson (1987) noted that for small $b$, $C_1(n), \ldots, C_b(n)$ are approximately independent Poisson-distributed random variables. The results of this paper may be viewed as confirmation and quantification of this observation. These total variation bounds may be used to prove a functional central limit theorem

TABLE 1
*Exact and estimated total variation distances*

| | | $\theta = 0.1$ | $\theta = 0.9$ | $\theta = 1.1$ | $\theta = 2.0$ |
|---|---|---|---|---|---|
| $n = 10$ | $b = 1$ | $8.85^*_{-3}\,(1.19_{-2})$ | $3.51_{-3}\,(1.58_{-1})$ | $3.96_{-3}\,(2.07_{-1})$ | $4.92_{-2}\,(4.85_{-1})$ |
| | $b = 2$ | $1.76_{-2}\,(2.64_{-2})$ | $1.30_{-2}\,(3.34_{-1})$ | $1.90_{-2}\,(4.16_{-1})$ | $1.19_{-1}\,(1.00_{-0})$ |
| | $b = 3$ | $2.67_{-2}\,(4.48_{-2})$ | $8.98_{-2}\,(5.37_{-1})$ | $1.24_{-1}\,(6.24_{-1})$ | $3.51_{-1}\,(> 1)$ |
| $n = 50$ | $b = 1$ | $1.66_{-3}\,(2.23_{-3})$ | $7.26_{-4}\,(3.36_{-2})$ | $8.03_{-4}\,(4.52_{-2})$ | $1.06_{-2}\,(1.15_{-1})$ |
| | $b = 3$ | $4.64_{-3}\,(6.96_{-3})$ | $1.86_{-3}\,(1.03_{-1})$ | $2.05_{-3}\,(1.36_{-1})$ | $2.69_{-2}\,(3.46_{-1})$ |
| | $b = 7$ | $1.04_{-2}\,(1.76_{-2})$ | $4.06_{-3}\,(2.52_{-1})$ | $4.46_{-3}\,(3.16_{-1})$ | $5.84_{-2}\,(8.08_{-1})$ |
| $n = 100$ | $b = 1$ | $8.21_{-4}\,(1.11_{-3})$ | $3.65_{-4}\,(1.70_{-2})$ | $4.02_{-4}\,(2.28_{-2})$ | $5.36_{-3}\,(5.86_{-2})$ |
| | $b = 4$ | $2.98_{-3}\,(4.56_{-3})$ | $1.20_{-3}\,(6.89_{-2})$ | $1.32_{-3}\,(9.14_{-2})$ | $1.76_{-2}\,(2.35_{-1})$ |
| | $b = 10$ | $7.05_{-3}\,(1.21_{-2})$ | $2.83_{-3}\,(1.78_{-1})$ | $3.12_{-3}\,(2.29_{-1})$ | $4.13_{-2}\,(5.88_{-1})$ |
| $n = 250$ | $b = 1$ | $3.27_{-4}\,(4.41_{-4})$ | $1.46_{-4}\,(6.82_{-3})$ | $1.61_{-4}\,(9.20_{-3})$ | $2.16_{-3}\,(2.38_{-2})$ |
| | $b = 6$ | $1.71_{-3}\,(2.70_{-3})$ | $6.94_{-4}\,(4.13_{-2})$ | $7.66_{-4}\,(5.52_{-2})$ | $1.03_{-2}\,(1.43_{-1})$ |
| | $b = 16$ | $4.39_{-3}\,(7.47_{-3})$ | $1.78_{-3}\,(1.13_{-1})$ | $1.96_{-3}\,(1.47_{-1})$ | $2.62_{-2}\,(3.81_{-1})$ |

Table gives exact $d_b(n)$ from (36), and (in parentheses) the smaller of the upper bounds (22) and (23).
*The notation $a_b$ means $a \times 10^b$.

for the Ewens sampling formula, proved by other methods by DeLaurentis and Pittel (1985) in the case $\theta = 1$, and Hansen (1990) and Donnelly, Kurtz and Tavaré (1991) for any $\theta > 0$. For further applications, see Arratia and Tavaré (1992b).

**Acknowledgments.** We would like to thank Frank Stenger for teaching us the utility of Darboux's method, and Jim Pitman and a referee for suggestions that improved the presentation of the paper.

## REFERENCES

ALDOUS, D. J. (1985). Exchangeability and related topics. *Lecture Notes in Math.* **1117** 1–198. Springer, New York.

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). Process approximations for the large components of random combinatorial structures. Unpublished manuscript.

ARRATIA, R. and TAVARÉ, S. (1992a). The cycle structure of random permutations. *Ann. Probab.* **20** 1567–1591.

ARRATIA, R. and TAVARÉ, S. (1992b). Limit theorems for combinatorial structures via discrete process approximations. *Random Structures and Algorithms*. To appear.

BARBOUR, A. D. (1990). Comment on "Poisson approximation and the Chen–Stein method" by R. Arratia, L. Goldstein and L. Gordon. *Statist. Sci.* **5** 425–427.

BARBOUR, A. D. (1992). Refined approximations for the Ewens sampling formula. *Random Structures and Algorithms*. To appear.

BLACKWELL, D. and MACQUEEN, J. B. (1973). Ferguson distributions via Pólya urn schemes. *Ann. Statist.* **1** 353–355.

BOLLOBÁS, B. (1985). *Random Graphs*. Academic, New York.

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

Diaconis, P. and Pitman, J. W. (1986). Unpublished lecture notes. Dept. Statist., Univ. California, Berkeley.

Donnelly, P. (1986). Partition structures, Pólya urns, the Ewens sampling formula and the ages of alleles. *Theoret. Population Biol.* **30** 271–288.

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.

Donnelly, P. and Tavaré, S. (1986). The ages of alleles and a coalescent. *Adv. in Appl. Probab.* **18** 1–19.

Erdös, P. and Turán, P. (1967). On some problems of statistical group theory. III. *Acta Math. Acad. Sci. Hungar.* **18** 309–320.

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theoret. Population Biol.* **3** 87–112.

Feller, W. (1945). The fundamental limit theorems in probability. *Bull. Amer. Math. Soc.* **51** 800–832.

Goldie, C. M. (1989). Records, permutations and greatest convex minorants. *Math. Proc. Cambridge Philos. Soc.* **106** 169–177.

Goncharov, V. L. (1944). Some facts from combinatorics. *Izv. Akad. Nauk SSSR Ser. Mat.* **8** 3–48. (See also: On the field of combinatory analysis. *Trans. Amer. Math. Soc.* **19** 1–46.)

Hansen, J. C. (1990). A functional central limit theorem for the Ewens sampling formula. *J. Appl. Probab.* **27** 28–43.

Hearn, A. C. (1987). REDUCE-3 User's Manual, Version 3.3. Rand Corporation Publication CP78.

Hoppe, F. M. (1984). Pólya-like urns and the Ewens sampling formula. *J. Math. Biol.* **20** 91–94.

Hoppe, F. M. (1987). The sampling theory of neutral alleles and an urn model in population genetics. *J. Math. Biol.* **25** 123–160.

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

Joyce, P. J. and Tavaré, S. (1987). Cycles, permutations and the structure of the Yule process with immigration. *Stochastic Process. Appl.* **25** 309–314.

Kelly, F. P. (1977). Exact results for the Moran neutral alleles model. *Adv. in Appl. Probab.* **9** 197–201.

Kolchin, V. F. (1971). A problem of the allocation of particles in cells and cycles of random permutations. *Theory Probab. Appl.* **16** 74–90.

Tavaré, S. (1987). The birth process with immigration, and the genealogical structure of large populations. *J. Math. Biol.* **25** 161–168.

Watterson, G. A. (1974). Models for the logarithmic species abundance distributions. *Theoret. Population Biol.* **6** 217–250.

Watterson, G. A. (1987). Estimating the proportion of neutral mutants. *Genetics Res. Cambridge* **501** 155–163.

Wilf, H. S. (1990). *Generatingfunctionology*. Academic, San Diego.

Richard Arratia
Department of Mathematics
University of Southern California
Los Angeles, California 90089-1113

A. D. Barbour
Institut für Angewandte Mathematik
Universität Zürich
Rämistrasse 74
CH-8001, Zürich
Switzerland

Simon Tavaré
Department of Mathematics
University of Southern California
Los Angeles, California 90089-1113