ON THE FUNCTIONAL CENTRAL LIMIT THEOREM FOR
THE EWENS SAMPLING FORMULA

BY PETER DONNELLY, THOMAS G. KURTZ ¹ AND SIMON TAVARE ²

Queen Mary and Westfield College, University of Wisconsin, Madison,
and University of Southern California

The Ewens sampling formula arises in population genetics and the
study of random permutations as a probability distribution on the set of
partitions (by allelic type in a sample, or according to cycle structure,
respectively) of the integer n for each n. It may be embedded naturally in
the familiar linear birth process with immigration. One consequence of this
is another proof of the functional central limit theorem for the Ewens
sampling formula.

1. Introduction. For each natural number n, the Ewens sampling for-
mula defines a distribution on the set \( \Pi_n \) of partitions of n. It is conventional
to represent such a partition in the form \( \pi = 1^{\alpha_1}2^{\alpha_2} \cdots n^{\alpha_n} \), where \( \alpha_i \) is the
number of parts of size \( i, i = 1, 2, \ldots, n \). The Ewens sampling formula with
parameter \( \theta > 0 \) assigns probability

\[
\frac{n!}{\theta_{(n)}} \prod_{i=1}^{n} \frac{\theta^{\alpha_i}}{i^{\alpha_i} \alpha_i !}
\]

(1.1)
to the partition \( 1^{\alpha_1}2^{\alpha_2} \cdots n^{\alpha_n} \) of n, where we have written \( \theta_{(n)} = \theta(\theta + 1) \)
\( \cdots (\theta + n - 1) \) for the ascending factorial of \( \theta \).

The distribution (1.1) originally arose in genetics [Ewens (1972)] as a robust
description of the random partition (according to the allelic type) of a sample of
n genes taken from a stationary, neutral, infinitely many alleles model. Note
that it also arises, for \( \theta = 1 \), as the description of the cycle lengths of a
uniformly distributed random permutation of n objects.

In many settings, interest centers on aspects of the distribution (1.1) for
large n. Our particular concern here is with the number of parts of the
partition of various sizes. Specifically, define the (step) functions \( K_n : [0, 1] \times \Pi_n \to \mathbb{R} \) by

\[
K_n(u, \pi) = \text{the number of parts in } \pi \text{ of length at most } n^u
\]

for \( \pi \in \Pi_n \). Our purpose here is to use a Poisson embedding argument to
prove the following functional central limit theorem for \( K_n \), due originally to

---

¹ Received April 1990; revised January 1991.
² Research supported in part by SERC Grant GR/F 59148 and NSF Grant DMS-89-01464.
   ² Research supported by SERC Grant GR/E 43898 and NSF Grants DMS-88-03284 and
   DMS-90-05833.
AMS 1980 subject classifications. 60C05, 60F17, 60J85, 92D10.
Key words and phrases. Random partitions, random permutations, Brownian motion.
THEOREM 1.1 [Hansen (1990)]. If \( \tilde{\pi}_n \) has distribution given by the Ewens sampling formula (1.1), then as \( n \to \infty \) the random element

\[
Y_n(u, \tilde{\pi}_n) = \frac{K_n(u, \tilde{\pi}_n) - \theta u \log n}{\sqrt{\theta \log n}}
\]

of \( D[0, 1] \) converges weakly to Brownian motion on \([0, 1]\).

REMARK 1. As Hansen notes, the theorem contains the central limit theorem for the number of parts of a partition with distribution (1.1), and, when \( \theta = 1 \), a functional central limit theorem for the cycle lengths of a random permutation, due originally to DeLaurentis and Pittel (1985).

REMARK 2. There are at least two other limiting regimes which are of interest for the Ewens sampling formula. For completeness we note that when normalized by \( n \), the sizes of the parts, written in decreasing order, of a partition with distribution (1.1) have a distribution which converges to the Poisson–Dirichlet distribution with parameter \( \theta \) [Kingman (1977, 1978)]. Other labelings of the parts also have nontrivial limiting distributions when normalized by \( n \). Also, if \( \tilde{\pi}_n \) has distribution (1.1), the joint distribution of the number of parts of \( \tilde{\pi}_n \) of sizes 1, 2, \ldots, \( k \), for any fixed \( k \), converges to that of independent Poisson random variables with means \( \theta, \theta/2, \ldots, \theta/k \), respectively.

2. The embedding. Recall the (linear) birth process with immigration, which we denote by \( \{I(t); t \geq 0\} \): at the points \( T_1 < T_2 < \cdots \) of a homogeneous Poisson process \( \mathcal{P}(\cdot) \) on \([0, \infty)\) of rate \( \theta \), individuals ("migrants") arrive and initiate families, each of which grows (independently of all other events) as a linear birth process of rate 1. If \( X_k(\cdot) \) denotes the birth process started by the \( k \)th immigrant, then \( X_k(0) = 1 \), and for \( i \neq j \),

\[
\lim_{h \downarrow 0} h^{-1} \mathbb{P}(X_k(t + h) = j | X_k(t) = i) = \begin{cases} i, & \text{if } j = i + 1 \\ 0, & \text{otherwise} \end{cases}
\]

for \( k = 1, 2, \ldots \). The process \( I(t) \) counts the total number of individuals (migrants and their offspring) present at time \( t \), with \( I(0) = 0 \):

\[
I(t) = \sum_{k=1}^{\mathcal{P}(t)} X_k(t - T_k).
\]

For \( n = 1, 2, \ldots \), let

\[
\tau_n = \inf\{t: I(t) = n\}
\]

be the time at which the birth and immigration process first reaches \( n \). It follows (e.g., from the representation of \( \tau_n \) as the sum of \( n \) independent exponential random variables with parameters \( \theta, \theta + 1, \ldots, \theta + n - 1 \)) that

\[
\tau_n / \log n \to 1 \quad \text{a.s. as } n \to \infty.
\]
Now, for $i = 1, 2, \ldots$, let

$$\alpha_i(t) = \# \{ k : T_k \leq t, \ X_k(t - T_k) = i \}$$

denote the number of families of size $i$ in the birth and immigration process at $t$. For our purposes, the crucial embedding property is the fact that when there is a total of $n$ individuals present, the family sizes form a partition that has distribution given by the Ewens sampling formula with parameter $\theta$ [Tavaré (1987)]. That is, $(\alpha_1(\tau_n), \alpha_2(\tau_n), \ldots, \alpha_n(\tau_n))$ has distribution (1.1).

The intuition behind the proof is the following. For large $n$, $\tau_n$ is close to $\log n$. Each family in the birth and immigration process grows exponentially, so that if we pretend that $\tau_n$ actually equals $\log n$, most of those families founded since time $(1 - u) \log n$ will have size less than $n^u$ at time $\log n$ (or $\tau_n$) and most of those families founded before time $(1 - u) \log n$ will have size larger than $n^u$ at time $\log n$. Thus $K_n(u)$ is like the number of families founded between time $(1 - u) \log n$ and time $\log n$. This is just the number of events in the underlying Poisson process of migrations over a time period of length $u \log n$. If we replace $K_n(\cdot)$ by $P(\log n) - P((1 - \cdot) \log n)$ in (1.2), then the theorem is just the functional central limit theorem for this Poisson process. Other applications of branching process embeddings in the study of urn schemes may be found, for example, in Athreya and Karlin (1968), Athreya and Ney (1972) and Holst, Kennedy and Quine (1988).

3. Proof of the theorem. Throughout, $\delta$ is fixed with $0 < \delta < 1/2$. For our construction,

$$K_n(u) = K_n(u, \hat{\tau}_n) = \sum_{k=1}^{\mathcal{P}(\tau_n)} 1 \{ X_k(\tau_n - T_k) \leq e^u \log n \}$$

(3.1) \[ \leq \sum_{k=1}^{\mathcal{P}(\tau_n)} 1 \{ X_k(\tau_n - T_k) \leq e^u \log n, \ X_k(\tau_n - T_k) \geq e^{\tau_n - T_k - (\log n)^{\delta}} \} \]

$$+ \sum_{k=1}^{\mathcal{P}(\tau_n)} 1 \{ X_k(\tau_n - T_k) < e^{\tau_n - T_k - (\log n)^{\delta}} \}.$$  

Denote the second term on the right of (3.1) by $R_1(n)$ and note that it does not depend on $u$. The first term is bounded above by

$$\sum_{k=1}^{\mathcal{P}(\tau_n)} 1 \{ e^{\tau_n - T_k - (\log n)^{\delta}} \leq e^u \log n \}

= \sum_{k=1}^{\mathcal{P}(\tau_n)} 1 \{ T_k \geq \tau_n - u \log n - (\log n)^{\delta} \}

= \mathcal{P}(\tau_n) - \mathcal{P}(\tau_n - u \log n - (\log n)^{\delta}).$$

(We adopt the convention that $\mathcal{P}(t) = 0$ if $t < 0$.)
In the other direction,
\[
K_n(u) = \mathcal{P}(\tau_n) - \sum_{k=1}^{\mathcal{P}(\tau_n)} 1\{X_k(\tau_n - T_k) > e^{u \log n}\}
\]
\[
(3.2) \geq \mathcal{P}(\tau_n) - \sum_{k=1}^{\mathcal{P}(\tau_n)} 1\{X_k(\tau_n - T_k) > e^{\tau_n - T_k + (\log n)^{\delta}}\}
\]
\[
- \sum_{k=1}^{\mathcal{P}(\tau_n)} 1\{X_k(\tau_n - T_k) > e^{u \log n}, X_k(\tau_n - T_k) \leq e^{\tau_n - T_k + (\log n)^{\delta}}\}.
\]
Denote the second term on the right of (3.2) by \(R_2(n)\), and again note that it does not depend on \(u\). Arguing as above then gives
\[
K_n(u) \geq \mathcal{P}(\tau_n) - \mathcal{P}(\tau_n - u \log n + (\log n)^{\delta}) - R_2(n).
\]
Thus, for all \(u \in [0, 1]\),
\[
\frac{\mathcal{P}(\tau_n) - \mathcal{P}(\tau_n - u \log n + (\log n)^{\delta}) - \theta u \log n}{\sqrt{\theta \log n}} - \frac{R_2(n)}{\sqrt{\theta \log n}} \leq Y_n(u) \equiv Y_n(u, \hat{\tau}_n)
\]
\[
\leq \frac{\mathcal{P}(\tau_n) - \mathcal{P}(\tau_n - u \log n - (\log n)^{\delta}) - \theta u \log n}{\sqrt{\theta \log n}} + \frac{R_1(n)}{\sqrt{\theta \log n}}.
\]
For \(0 \leq s < \infty\), define the random element \(W_n(\cdot)\) of \(D[0, \infty)\) by
\[
W_n(s) = \frac{\mathcal{P}(s \log n) - \theta s \log n}{\sqrt{\theta \log n}}
\]
and note that we can write (3.3) as
\[
W_n\left(\frac{\tau_n}{\log n}\right) - W_n\left(\frac{\tau_n}{\log n} + (\log n)^{\delta - 1} - u\right)
\]
\[
- \sqrt{\theta (\log n)^{\delta - 1/2}} - \frac{R_2(n)}{\sqrt{\theta \log n}}
\]
\[
(3.4) \leq Y_n(u)
\]
\[
\leq W_n\left(\frac{\tau_n}{\log n}\right) - W_n\left(\frac{\tau_n}{\log n} - (\log n)^{\delta - 1} - u\right)
\]
\[
+ \sqrt{\theta (\log n)^{\delta - 1/2}} + \frac{R_1(n)}{\sqrt{\theta \log n}}.
\]
Thus
\[
\sup_{0 \leq u \leq 1} \left| Y_n(u) - (W_n(1) - W_n(1 - u)) \right| \\
\leq \frac{R_1(n) + R_2(n)}{\sqrt{\theta \log n}} + \sqrt{\theta} (\log n)^{\delta - 1/2}
\]
\[
+ \sup_{0 \leq u \leq 1} \left| W_n \left( \frac{\tau_n}{\log n} - (\log n)^{\delta - 1} - u \right) - W_n(1 - u) \right| \\
+ \sup_{0 \leq u \leq 1} \left| W_n \left( \frac{\tau_n}{\log n} + (\log n)^{\delta - 1} - u \right) - W_n(1 - u) \right| \\
+ \left| W_n \left( \frac{\tau_n}{\log n} \right) - W_n(1) \right|.
\]

The Donsker invariance principle gives the convergence in distribution of $W_n$ to standard Brownian motion $W$. To see the implication of this convergence more easily, the Skorohod representation theorem [Ethier and Kurtz (1986), Theorem 3.1.8] says that we may "pretend" that $\sup_{0 \leq u \leq 1} |W_n(u) - W(u)| \to 0$ a.s. Since $\tau_n/\log n \to 1$ a.s. and $W$ is continuous, we see that the last three terms on the right of the above inequality tend to 0 in probability. We will show below that
\[(3.5) \quad R_1(n) \to_p 0, \quad R_2(n) \to_p 0 \quad \text{as } n \to \infty, \]
and hence conclude that
\[
\sup_{0 \leq u \leq 1} \left| Y_n(u) - (W_n(1) - W_n(1 - u)) \right| \to_p 0.
\]
From the Skorohod representation, we see immediately that $Y_n(\cdot) \Rightarrow W(1) - W(1 - \cdot)$, which gives the desired result.

It remains to establish (3.5). We will use the following lemma.

**Lemma 3.1.** Suppose that $\{M(t), t \geq 0\}$ is a right-continuous, uniformly integrable nonnegative martingale with respect to a filtration $\{\mathcal{F}_t, t \geq 0\}$. Let $\lim_{t \to \infty} M(t) = M$ a.s. Then for $\gamma > 0$,
\[
\mathbb{P} \left( \inf_{t \geq 0} M(t) < \gamma \right) \leq \mathbb{P}(M \leq \sqrt{\gamma}) + \sqrt{\gamma}.
\]

**Proof.** We may write
\[(3.7) \quad M(t) = \mathbb{E}(M | \mathcal{F}_t) \quad \text{a.s.}\]
For fixed $\gamma > 0$, define the stopping time $\sigma = \inf(t \geq 0: M(t) < \gamma)$. For any $t > 0$ we have

\[
\mathbb{P}(M \leq \sqrt{\gamma}) \geq \mathbb{E}(\mathbb{P}(M \leq \sqrt{\gamma} | \mathcal{F}_{t \wedge \sigma}) 1\{\sigma < t\}) \\
= \mathbb{P}(\sigma < t) - \mathbb{E}(\mathbb{P}(M > \sqrt{\gamma} | \mathcal{F}_{t \wedge \sigma}) 1\{\sigma < t\}) \\
\geq \mathbb{P}(\sigma < t) - \mathbb{E}(\mathbb{P}(M | \mathcal{F}_{t \wedge \sigma}) 1\{\sigma < t\})/\sqrt{\gamma} \\
\geq \mathbb{P}(\sigma < t) - \sqrt{\gamma},
\]

using Markov's inequality, the representation (3.7), and the fact that by the definition of $\sigma$, $\mathcal{M}(t \wedge \sigma)1(\sigma < t) \leq \gamma$. Thus

$$\mathbb{P}\left( \inf_{t \geq 0} \mathcal{M}(t) < \gamma \right) = \mathbb{P}(\sigma < \infty) \leq \mathbb{P}(\mathcal{M} \leq \sqrt{\gamma}) + \sqrt{\gamma},$$

completing the proof. □

If $\{X(t); t \geq 0\}$ is a rate 1 linear birth process, then it is well known that $\mathcal{M}(t) = e^{-t}X(t)$ is an $L_2$ bounded martingale with respect to the filtration $\{\mathcal{F}_t, t \geq 0\}$ generated by $X(\cdot)$. See, for example, Athreya and Ney (1972, page 111). Further, $\mathcal{M}$ is a mean-one exponential random variable, so we may use (3.6) to see that

$$\mathbb{P}\left( \inf_{t \geq 0} \mathcal{M}(t) < \gamma \right) \leq 2\sqrt{\gamma}.$$

REMARK. We thank J. W. Pitman for pointing out to us that the argument of Dubins and Gilat (1978) may be used to show that if $\mathbb{P}(\mathcal{M} \leq \gamma) \leq B \gamma$ for $0 \leq \gamma \leq \gamma_0$, then $\mathbb{P}(\inf_{t \geq 0} \mathcal{M}(t) \leq \gamma) \leq 2B \gamma$ for $0 \leq \gamma \leq \gamma_0/2$. In particular, this strengthens the right-hand side of the inequality (3.9) to $2\gamma$.

**LEMMA 3.2.** As $n \to \infty$,

$$R_1(n) = \sum_{k=1}^{\mathcal{P}(\tau_n)}1\left\{X_k(\tau_n - T_k) < e^{\tau_n - T_k - (\log n)^{\beta}}\right\} \to_{\mathbb{P}} 0$$

and

$$R_2(n) = \sum_{k=1}^{\mathcal{P}(\tau_n)}1\left\{X_k(\tau_n - T_k) > e^{\tau_n - T_k + (\log n)^{\beta}}\right\} \to_{\mathbb{P}} 0.$$

**PROOF.** Fix $c > 1$ and adopt the convention that for $k = 1, 2, \ldots, X_k(t) = 1$ if $t < 0$. Then

$$R_1(n) \leq \sum_{k=1}^{\mathcal{P}(c \log n)}1\left\{e^{-(\tau_n - T_k)}X_k(\tau_n - T_k) < e^{-(\log n)^{\beta}}\right\}$$

$$+ \mathcal{P}(\tau_n)1(\tau_n > c \log n).$$

It follows from the fact that $\tau_n/\log n \to 1$ a.s. that the second term on the right of (3.12) converges to 0 almost surely. The first term is nonnegative and bounded above by

$$\mathcal{P}(c \log n) \sum_{k=1}^{\mathcal{P}(c \log n)}1\{e^{-t}X_k(t) < e^{-(\log n)^{\beta}}\} \text{ for some } t \geq 0,$$

but (3.9) and the independence of the Poisson process $\mathcal{P}(\cdot)$ and the birth processes $X_k(\cdot), k = 1, 2, \ldots$, ensure that the mean of this random variable converges to 0 as $n \to \infty$, so that in particular it converges to 0 in probability, which establishes (3.10).
To establish (3.11), note that

\[
R_2(n) \leq \sum_{k=1}^{\mathcal{P}(c \log n)} 1\{e^{-(\tau_n - T_k)} \wedge 0 X_k(\tau_n - T_k) > e^{(\log n)^p}\} + \mathcal{P}(\tau_n)I\{\tau_n > c \log n\}.
\]

As above, the second term converges to 0 almost surely, and the first term is nonnegative with mean bounded above by

\[
c\theta \log n \mathbb{P}\left( \sup_{t>0} e^{-t}X_1(t) > e^{(\log n)^p} \right) \leq c\theta \log n e^{-(\log n)^p} \to 0 \quad \text{as } n \to \infty
\]

by Doob’s inequality. □

REFERENCES


