Limit Theorems for Combinatorial Structures via Discrete Process Approximations

Richard Arratia and Simon Tavaré*
Department of Mathematics, University of Southern California, Los Angeles, CA 90089–1113

ABSTRACT

Discrete functional limit theorems, which give independent process approximations for the joint distribution of the component structure of combinatorial objects such as permutations and mappings, have recently become available. In this article, we demonstrate the power of these theorems to provide elementary proofs of a variety of new and old limit theorems, including results previously proved by complicated analytical methods. Among the examples we treat are Brownian motion limit theorems for the cycle counts of a random permutation or the component counts of a random mapping, a Poisson limit law for the core of a random mapping, a generalization of the Erdös–Turán Law for the log-order of a random permutation and the smallest component size of a random permutation, approximations to the joint laws of the smallest cycle sizes of a random mapping, and a limit distribution for the difference between the total number of cycles and the number of distinct cycle sizes in a random permutation. © 1992 John Wiley & Sons, Inc.

Key Words: random mappings, random permutations, functional limit theorem, Erdös–Turán law, Poisson processes

1. INTRODUCTION

Many random combinatorial structures may be described in the following broad terms: for each natural number n, let $C_1(n), C_2(n), \ldots, C_n(n)$ be the number of

* The authors were supported in part by NSF grant DMS 90-05833 (R.A.A., S.T.) and NIH grant GM 41746 (S.T.). We thank Andrew Barbour for helpful comments on earlier drafts of this article.

Random Structures and Algorithms, Vol. 3, No. 3 (1992)
© 1992 John Wiley & Sons, Inc. CCC 1042–9832/92/030321–25$04.00

components of sizes 1, 2, . . . , n in the structure. For large n, these dependent counts $C_j(n)$ may be approximated by a limit process on $\mathbb{N} \equiv \{1, 2, \ldots\}$, in the sense that as $n \to \infty$

$$(C_1(n), C_2(n), \ldots) \Rightarrow (Z_1, Z_2, \ldots) \tag{1}$$

where the Z_i, $i = 1, 2, \ldots$ are independent random variables, and $\Rightarrow$ denotes convergence in distribution. In the case of a uniform random permutation, in which components are cycles, the Z_i are Poisson distributed with mean

$$\mathbb{E}(Z_i) = \frac{1}{i}, \tag{2}$$

as shown by Goncharov [25] and Kolchin [29]. In the case of a random mapping function, uniformly chosen from the n^n possibilities, the Z_i are Poisson distributed with mean

$$\mathbb{E}(Z_i) = \frac{1}{i} \sum_{r=0}^{i-1} \frac{e^{-i} i^r}{r!}, \tag{3}$$

as shown by Kolchin [30].

Limit distributions other than the Poisson may arise, a common feature being the existence of a parameter $\theta > 0$ such that $\mathbb{E} i Z_i \to \theta$ and $\mathbb{P}(Z_i = 1) \sim \theta / i$ as $i \to \infty$. A description of these limits in general is given in Arratia and Tavaré [5]. For example, the case in which Z_i has the negative binomial distribution with parameters $N_q(i)$ and q^{-i} where $q \in \mathbb{N}$ is fixed and

$$N_q(i) = \frac{1}{i} \sum_{d \mid i} \mu(i/d) q^d,$$

arises in the study of necklaces (Metropolis and Rota [34, 35]), card shuffling (Diaconis, McGrath, and Pitman [42]), and in factorization of monic polynomials over a finite field (cf. Lidl and Neiderreiter [33, p. 84]). Further details may be found in Arratia, Barbour, and Tavaré [7].

For most purposes (1) is not strong enough to imply that natural properties of the combinatorial object can be derived from the limiting independent process. This is because (1) only involves convergence of the distribution of $(C_1(n), \ldots, C_b(n))$ for each fixed b as $n \to \infty$. Many natural properties depend jointly on all component counts, albeit only weakly on the largest ones. Estimates are needed in which b and n grow simultaneously. There are now explicit estimates on the behavior of the total variation distance $d_b(n)$ between the law of $(C_1(n), \ldots, C_b(n))$ and the law of $(Z_1, \ldots, Z_b)$ as a function of b and n. Such estimates allow the small cycle sizes, of order up to $b = o(n)$, to be decoupled into independent random variables, with an upper bound on the error involved. It is the purpose of this article to show how this decoupling may be used to unify and simplify the proofs of limit theorems for a variety of functionals of certain random combinatorial structures. The basic strategy is as follows. For an appropriate choice of b with $b \to \infty$, $b/n \to 0$, the components of size greater than b make a negligible contribution to the functional, and this can often be shown easily by Chebychev-type inequalities. The components of size at most b are approximated by the limit process, the error being controlled by the bound on the total variation distance $d_b(n)$. The functional evaluated at the limit process is easily analyzed using independence. In this article, we illustrate this approach for the examples of random permutations and random mappings.

Here is an outline of the article. Section 2 gives examples which correspond to linear functionals of the cycle structure of random permutations. Section 3 treats the Erdös–Turán law for the group-order of a permutation, and Section 4 discusses nonlinear functions. Section 5 shows how similar results are proved for the component structure of random mappings. These first five sections give consequences of "naive" Poisson process approximations; they exploit convergence to zero of total variation distance, without using the available bounds. Section 6 gives an example of the additional power supplied by the uniformity of these bounds in studying the cycle structure of a random mapping.

Arratia, Goldstein, and Gordon [2] treat the example of cycles in random graphs using a Poisson process approximation, emphasizing how the Chen–Stein method yields bounds on the total variation distance for processes, and giving one example of the application of this approximation, in the spirit of Theorem 4 below. See also Barbour, Holst, and Janson [43] which treats many other combinatorial examples. Flajolet and Soria [24] discuss Gaussian limit laws for combinatorial structures using generating function methods. Other recent approaches to random mappings are described in Kolchin [32], Flajolet and Odlyzko [23], and Aldous and Pitman [1].

A. Total Variation Distance

We end the introduction by recalling some standard facts about total variation distance. For 1ə$b \leq n$, let $d_b(n)$ be the total variation distance between the law of $\mathbf{C}_b(n) \equiv (C_1(n), \ldots, C_b(n))$ and the law of $\mathbf{Z}_b \equiv (Z_1, \ldots, Z_b)$:

$$\begin{aligned} d_b(n) &\equiv \|\mathscr{L}(C_b(n)) - \mathscr{L}(Z_b)\| \\ &= \sup_{A \subseteq \mathbb{Z}_+^b} |\mathbb{P}(C_b(n) \in A) - \mathbb{P}(Z_b \in A)| , \end{aligned} \tag{4}$$

where $\mathbb{Z}_+ \equiv \{0, 1, \ldots\}$. An equivalent definition of $d_b(n)$ is

$$d_b(n) = \frac{1}{2} \sum_{\mathbf{a} \in \mathbb{Z}_+^b} |\mathbb{P}(\mathbf{C}_b(n) = \mathbf{a}) - \mathbb{P}(\mathbf{Z}_b = \mathbf{a})| . \tag{5}$$

Further,

$$d_b(n) = \inf \mathbb{P}(\mathbf{C}_b(n) \neq \mathbf{Z}_b) , \tag{6}$$

the infimum being taken over all couplings of $\mathbf{C}_b(n)$ and $\mathbf{Z}_b$ on the same probability space. There are maximal couplings that attain this bound.

2. PERMUTATIONS

We will discuss random permutations in a one-parameter setting which includes the usual uniform distribution as a special case. The Ewens sampling formula with parameter $\theta > 0$ may be thought of as the measure on the permutations of $\{1, 2, \ldots, n\}$ whose density with respect to uniform measure is proportional to θ^k, where k is the number of cycles in the permutation. The special case $\theta = 1$ corresponds to uniform measure. The set of all permutations with cycle index $(a_1, a_2, \ldots, a_n)$ (that is, having a_j cycles of length j for $j = 1, \ldots, n$) has probability

$$P_n(a_1, \ldots, a_n) = \frac{n!}{\theta_{(n)}} \prod_{j=1}^{n} \left(\frac{\theta}{j}\right)^{a_j} \frac{1}{a_j!} 1\left\{\sum_{j=1}^{n} j a_j = n\right\}, \tag{7}$$

where we have denoted rising factorials by

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

This formula was derived by Ewens [21] in the context of population genetics, where a_j is the number of alleles represented by j genes in a sample of n genes taken from a large population; θ is a parameter that measures the mutation rate.

We let $C_j \equiv C_j(n)$ be the number of cycles of size j in an n-permutation, so that $C_j(n) \equiv 0$ if $j > n$. Under the Ewens sampling formula for fixed θ, $(C_1(n), C_2(n), \ldots) \Rightarrow (Z_1, Z_2, \ldots)$, where the Z_i are independent Poisson random variables with mean

$$\mathbb{E} Z_i = \frac{\theta}{i} \tag{8}$$

In fact it is possible to couple closely the cycle counting processes for all n, together with the limiting Poisson process, on a common probability space, as the following results from Arratia, Barbour, and Tavaré [6] show.

Theorem 1. *Let $\{\xi_j, j \geq 1\}$ be a sequence of independent Bernoulli random variables satisfying*

$$\mathbb{P}(\xi_j = 1) = \frac{\theta}{\theta + j - 1} . \tag{9}$$

For $j \leq n$, define

$$C_j(n) = \sum_{i=1}^{n-j} \xi_i (1 - \xi_{i+1}) \cdots (1 - \xi_{i+j-1}) \xi_{i+j} + \xi_{n-j+1} (1 - \xi_{n-j+2}) \cdots (1 - \xi_n) , \tag{10}$$

and for $j > n$ define $C_j(n) \equiv 0$. Define $C_j(\infty) \equiv Z_j$ by

$$Z_j = \sum_{i=1}^{\infty} \xi_i (1 - \xi_{i+1}) \cdots (1 - \xi_{i+j-1}) \xi_{i+j} . \tag{11}$$

Then $(C_1(n), \ldots, C_n(n))$ has the distribution (7), and the Z_j are independent Poisson random variables with $\mathbb{E} Z_j = \theta / j$. Further,

$$\sum_{j=1}^{n} C_j(n) = \sum_{j=1}^{n} \xi_j , \tag{12}$$

and for each j

$$C_j(n) \le Z_j + 1(J_n = j) , \tag{13}$$

where $J_n \in \{1, 2, \ldots, n\}$ *is defined by*

$$J_n = \min\{ j \ge 1 : \xi_{n-j+1} = 1\} \tag{14}$$

Finally, as $n \to \infty$

$$\sum_{j=1}^{n} \mathbb{E}|C_j(n) - Z_j| = O(1) . \tag{15}$$

Using this coupling, they proved *inter alia*

Theorem 2. *Let* $(C_1(n), C_2(n), \ldots)$ *be the cycle counting process for the Ewens sampling formula, and let* $(Z_1, Z_2, \ldots)$ *be the Poisson process on* $\mathbb{N}$ *determined by* (8). *For* $1 \le b \le n$, *let* $d_b(n)$, *defined in* (4), *be the total variation distance between* $(C_1(n), \ldots, C_b(n))$ *and* $(Z_1, \ldots, Z_b)$. *Then*

$$d_b(n) \to 0 \text{ if, and only if, } b = o(n) \tag{16}$$

The following result, which is useful in what follows, is an immediate consequence of (15):

Lemma 1. *There is a coupling of* $\{C_j(n), j \ge 1, n \ge 1\}$ *and* $\{Z_j, j \ge 1\}$ *such that*

$$R_n^* = \frac{\sum_{j=1}^{n} |C_j(n) - Z_j|}{\sqrt{\theta \log n}} . \tag{17}$$

converges in probability to 0 *as* $n \to \infty$.

Remark. In Lemma 1, the normalization by $\sqrt{\log n}$ may be replaced by any function of n tending to infinity with n.

A. The Number of Cycles

The first example sets the scene for the technique that will be employed throughout the article. Define

$$K_n = \sum_{j=1}^{n} C_j(n) ,$$

the number of cycles in a random n-permutation. From the representation (12), it follows that

$$\mathbb{E}K_n = \sum_{j=1}^{n} \frac{\theta}{\theta + j - 1}, \tag{18}$$

so that

$$\theta \log(n/\theta) \le \mathbb{E}K_n \le 1 + \theta + \theta \log n. \tag{19}$$

It is well known that K_n, appropriately centred and scaled, has asymptotically a standard Normal distribution:

Theorem 3. *As $n \to \infty$,*

$$\frac{K_n - \theta \log n}{\sqrt{\theta \log n}} \Rightarrow \mathrm{N}(0, 1). \tag{20}$$

Remark. This result has a long history. It is due originally to Goncharov [25] and there are now many different proofs. Feller [22] gives a representation of K_n as a sum of independent (but not identically distributed) Bernoulli random variables, Shepp and Lloyd [37] use generating functions, Kolchin [29] uses a representation in terms of random allocation of particles into cells. The authors above all considered the case $\theta = 1$, but their methods extend to general θ. In fact, Feller's proof uses the special case $\theta = 1$ of (12), and its generalization is simply the observation that $K_n = \sum_{j=1}^{n} \xi_j$ is asymptotically normal, via the Lindeberg–Feller conditions.

Remark. The results of Barbour and Hall [9] may be combined with the representation of K_n as a sum of n independent, nonidentically distributed Bernoulli random variables to show that if P_n is a Poisson random variable with mean $\mathbb{E}K_n$ given by (18), then

$$\|\mathscr{L}(K_n) - \mathscr{L}(P_n)\| \asymp \frac{1}{\log n},$$

a result that is stronger than Theorem 3.

Proof. The present proof is intended to serve as a model for the other proofs in this section. The idea is to write

$$\frac{K_n - \theta \log n}{\sqrt{\theta \log n}} = \frac{\sum_{j=1}^{n} Z_j - \theta \log n}{\sqrt{\theta \log n}} + R_n, \tag{21}$$

where the remainder term R_n is given by

$$R_n = \frac{\sum_{j=1}^{n} (C_j(n) - Z_j)}{\sqrt{\theta \log n}}.$$

The Z_j are independent Poisson random variables satisfying (6), so that $\sum_{j=1}^{n} Z_j$ has a Poisson distribution with mean and variance $\sum_{j\le n} \theta/j \sim \theta \log n$. It is now immediate that

$$\frac{\sum_{j=1}^{n} Z_j - \theta \log n}{\sqrt{\theta \log n}} \Rightarrow N(0, 1) .$$

Since $|R_n| \le R_n^*$ and, by Lemma 1, $R_n^* \to_p 0$, which is constant, the result follows from Slutsky's Theorem (Billingsley [11, p. 25]). ∎

B. Cycle Lengths Modulo r

In this section, we give an example that shows more fully the power of Theorem 2. Choose and fix any integer $r \ge 1$, and define h_n: $\mathbb{R}^n \to \mathbb{R}^r$ by

$$h_n(x_1, \ldots, x_n) = \Bigl(\sum_{j\le n;\, j = i (\mathrm{mod}\ r)} x_j,\ i = 0, 1, \ldots, r-1 \Bigr)$$

Observe that the sum of the r components of $h_n(\mathbf{C}_n)$ is K_n, so we are considering a refinement of K_n. Let μ_n be a constant r-vector with elements $\theta \log n/r$. We then have

Theorem 4. *As $n \to \infty$,*

$$(\theta \log n/r)^{-1/2}(h_n(\mathbf{C}_n) - \mu_n) \Rightarrow N_r(0, I) , \tag{22}$$

where $N_r(0, I)$ is the r-dimensional standard normal distribution with independent coordinates.

Proof. As in the proof of Theorem 3, the idea is to replace $\mathbf{C}_n$ in (22) by $\mathbf{Z}_n$, for which the stated result is elementary to prove. The error in this approximation is

$$R_n \equiv (\theta \log n/r)^{-1/2}(h_n(\mathbf{C}_n) - h_n(\mathbf{Z}_n)) .$$

But from (17) and Lemma 1 we see that

$$\begin{aligned} \|R_n\|_1 &\le (\theta \log n/r)^{-1/2} \sum_{j=1}^{n} |C_j(n) - Z_j| \\ &= \sqrt{r} R_n^* \to_p 0 , \end{aligned}$$

completing the proof. ∎

C. A Functional Central Limit Theorem

In this section, we provide an elementary proof of Hansen's [27] functional version of the central limit result (20). To this end, define a random element $Y_n(\cdot)$

of $D[0, 1]$ by

$$Y_n(t) = \frac{\sum_{j=1}^{\lfloor n^t \rfloor} C_j(n) - \theta t \log n}{\sqrt{\theta \log n}}, \quad 0 \leq t \leq 1.$$

Theorem 3 asserts that $Y_n(1) \Rightarrow N(0, 1)$ as $n \to \infty$. The functional version is

Theorem 5 (Hansen [27]). *As* $n \to \infty$,

$$Y_n(\cdot) \Rightarrow \text{standard Brownian motion on } [0, 1] \tag{23}$$

Remark. The special case $\theta = 1$ of Theorem 5 was proved first by DeLaurentis and Pittel [14]. Another approach to the general case is given in Donnelly, Kurtz, and Tavaré [17].

Proof. Define the process $\{W_n(t), 0 \leq t \leq 1\}$ by

$$W_n(t) = \frac{\sum_{j=1}^{\lfloor n^t \rfloor} Z_j - \theta t \log n}{\sqrt{\theta \log n}},$$

and let

$$R_n(t) = \frac{\sum_{j=1}^{\lfloor n^t \rfloor} (C_j(n) - Z_j)}{\sqrt{\theta \log n}},$$

so that

$$Y_n(t) = W_n(t) + R_n(t).$$

We will show that the functionals $W_n(\cdot)$ of the Poisson process converge weakly to Brownian motion and that $R_n(\cdot) \to_P 0$ in the sup norm.

To see that $W_n(\cdot)$ converges weakly to standard Brownian motion, define $s(0) = 0$, $s(j) = \theta(1 + 1/2 + \cdots + 1/j)$, $j \geq 1$, and let $\{\mathcal{P}(t), t \geq 0\}$ be a rate one Poisson process with $\mathcal{P}(0) = 0$. For $t > 0$, we have

$$\begin{aligned}\sum_{j=1}^{\lfloor n^t \rfloor} Z_j &\overset{\mathrm{d}}{=} \sum_{j=1}^{\lfloor n^t \rfloor} (\mathcal{P}(s(j)) - \mathcal{P}(s(j-1))) \\ &= \mathcal{P}(s(\lfloor n^t \rfloor)).\end{aligned} \tag{24}$$

The functional central limit theorem for the Poisson process (cf. Ethier and Kurtz [20, p. 263]) shows that $s(n)^{-1/2}(\mathcal{P}(s(\lfloor n^t \rfloor)) - s(\lfloor n^t \rfloor))$ converges weakly to Brownian motion on $[0, 1]$ starting from 0. The corresponding result for $W_n(\cdot)$ then follows because $s(n) \sim \theta \log n$ and $\sup_{0 \leq t \leq 1} |\theta t \log n - s(\lfloor n^t \rfloor)| \leq 1$.

To show that $\sup_{0\le t\le 1} |R_n(t)| \to_P 0$ we may use (17) and Lemma 1 once more:

$$\begin{aligned}\sup_{0\le t\le 1} |R_n(t)| &\le \sup_{0\le t\le 1} \sum_{j=1}^{\lfloor n^t \rfloor} |C_j(n) - Z_j| / \sqrt{\theta \log n} \\ &= \sum_{j=1}^{n} |C_j(n) - Z_j| / \sqrt{\theta \log n} \\ &\le R_n^* \to_P 0 ,\end{aligned}$$

completing the proof. ∎

D. Linear Combinations

The total variation estimates may be used to study the asymptotic behavior of other weighted averages of the cycle counting process, by comparing them to the same weighted average of the Poisson process $Z_1, Z_2, \ldots$. For example, we have the following result for linear combinations:

Theorem 6. *For* $1 \le b \le n$, *define* $S_b(n) = \Sigma_{j=1}^{b} w_{nj} C_j(n)$, *and let* $S_b^*(n) = \Sigma_{j=1}^{b} w_{nj} Z_j$. *Then*

$$\|\mathscr{L}(S_b(n)) - \mathscr{L}(S_b^*(n))\| \le d_b(n) .$$

Proof. This is a consequence of the fact that

$$\{S_b(n) \ne S_b^*(n)\} \subseteq \{(C_1(n), \ldots, C_b(n)) \ne (Z_1, \ldots, Z_b)\} ,$$

and the definition of $d_b(n)$. ∎

Some examples of limiting behavior for linear combinations in the case $\theta = 1$ may be found in Kolchin [32, p. 50 ff], for example.

E. The Smallest Cycles

In this section we analyze some aspects of the smallest cycle sizes. For any vector $\mathbf{a} \in \mathbb{Z}_+^\infty$ of cycle counts, let b_r be the functional that records the rth smallest cycle length:

$$\begin{aligned} b_r(\mathbf{a}) &= \inf\{ j : a_1 + \cdots + a_j \ge r \} , \quad r = 1, 2, \ldots \\ &= \infty, \text{if no such } j \end{aligned}$$

and let h_m be the functional that records the sizes of the m smallest cycles:

$$h_m(\mathbf{a}) = (b_1, \ldots, b_m) .$$

Let $\mathbf{C}(n) \equiv (C_1(n), C_2(n), \ldots, C_n(n), 0, 0, \ldots)$ be the cycle counting process. Then $b_r(\mathbf{C}(n))$ is the length of the rth smallest cycle, and $h_m(\mathbf{C}(n))$ is the process of the m smallest cycle lengths. If $\mathbf{Z} \equiv (Z_1, Z_2, \ldots)$, then $b_r(\mathbf{Z})$ and $h_m(\mathbf{Z})$ are the corresponding functionals for the Poisson process $\mathbf{Z}$ of counts.

The one-dimensional distributions are elementary to analyze, since $b_r(\mathbf{C}(n)) > j$ if, and only if, $C_1 + \cdots + C_j < r$. Hence

$$|\mathbb{P}(b_r(\mathbf{C}(n)) > j) - \mathbb{P}(b_r(\mathbf{Z}) > j)| = \left|\mathbb{P}\Big(\sum_{i \le j} C_i < r\Big) - \mathbb{P}\Big(\sum_{i \le j} Z_i < r\Big)\right| \le d_j(n) .$$

Since $Z_1 + \cdots + Z_j$ has a Poisson distribution with mean $\theta(1 + 1/2 + \cdots + 1/j)$, the distribution of $b_r(\mathbf{Z})$ is readily computed. This distribution is given in the case $\theta = 1$ by Shepp and Lloyd [37].

A process version of the result is contained in

Theorem 7. *The total variation distance*

$$d_m^* \equiv \|\mathscr{L}(h_m(\mathbf{C}(n))) - \mathscr{L}(h_m(\mathbf{Z}))\| \tag{25}$$

tends to 0 *if* $m \equiv m(n) \le (1 - \epsilon)\theta \log n$ *for fixed* $\epsilon > 0$. *In fact,* $d_m^* \to 0$ *if, and only if,* $\omega_n \equiv (\theta \log n - m)/\sqrt{\log n}$ *satisfies* $\omega_n \to \infty$.

Proof. For the necessity of the condition $\omega_n \to \infty$, note that under any coupling

$$\Big\{\sum_{i \le n} Z_i < m\Big\} \subseteq \{h_m(\mathbf{C}(n)) \ne h_m(\mathbf{Z})\} ,$$

so that $d_m^* \ge \mathbb{P}(\Sigma_{i \le n} Z_i < m)$, which, by the central limit theorem, tends to 0 iff $\omega_n \to \infty$.

For the sufficiency, observe that for any m and b

$$\{h_m(\mathbf{C}(n)) \ne h_m(\mathbf{Z})\} \subseteq \{(C_1, \ldots, C_b) \ne (Z_1, \ldots, Z_b)\} \cup \Big\{\sum_{i \le b} Z_i < m\Big\} .$$

It follows that

$$d_m^* \le d_b(n) + \mathbb{P}\Big(\sum_{i \le b} Z_i < m\Big) . \tag{26}$$

Now it is possible to choose b in such a way that

$$\mathbb{E}(Z_1 + \cdots + Z_b) = \theta \sum_{j \le b} 1/j = \theta \log n - \omega_n \sqrt{\log n}/2 + \delta_n ,$$

where $0 \le \delta_n < \theta$. It follows that

$$\frac{b}{n} \asymp \exp\Big(-\frac{\omega_n}{2\theta}\sqrt{\log n}\Big) e^{\delta_n} ,$$

so that $b/n \to 0$. Theorem 2 then shows that $d_b(n) \to 0$. In addition, the central limit theorem shows that for such a choice of b, $\mathbb{P}\Big(\sum_{i \le b} Z_i < m\Big) \to 0$, since $\mathrm{var}(Z_1 + \cdots + Z_b) = O(\log n)$. This completes the proof. ∎

Related material appears in Kolchin [32, p. 46 ff].

3. THE ERDÖS–TURÁN LAW

When $\theta = 1$, the distribution (7) corresponds to uniform measure on the symmetric group, S_n. Among the vast literature in this area is a beautiful result due to Erdös and Turán [19] concerning the asymptotic normality of the log of the order O_n of a randomly chosen element of S_n. Their proof is based on showing (Erdös and Turán [18]) that $\log O_n$ is relatively close to $\log P_n$, where $P_n \equiv \Pi_{j=1}^n j^{C_j(n)}$ is the product of the cycle lengths, and that $\log P_n$, suitably centred and scaled, is asymptotically normally distributed.

We give a relatively simple proof of the Erdös–Turán result. This proof extends the Erdös–Turán law to all θ. Our proof has three steps. First the coupling in Theorem 1 is used to show that $|\log P_n - \log O_n|$ is readily controlled by the corresponding functional of the Poisson process. The second step uses a moment calculation for the Poisson process to show that this functional of the Poisson process is negligible relative to $\log^{3/2} n$. The last step, which is similar to the method used to prove Theorems 3, 4, and 5, shows that $\log P_n$ is close to the corresponding functional of the Poisson process.

We begin with the following deterministic lemma. Let $\mathbf{a} \in \mathbb{Z}_+^n$, and define

$$r(\mathbf{a}) = \frac{\prod_{i \le n} i^{a_i}}{1\,\mathrm{cm}\{i : a_i > 0\}} .$$

Lemma 2. *For* $\mathbf{a}$, $\mathbf{b} \in \mathbb{Z}_+^n$, *and* $\mathbf{e}_j = (\delta_{ij}, i = 1, \dots, n)$, $j \le n$ *satisfying* $\mathbf{a} \le \mathbf{b} + \mathbf{e}_j$, *we have*

$$1 \le r(\mathbf{a}) \le nr(\mathbf{b}) . \tag{27}$$

Proof. The first inequality in (27) is immediate. To establish the second inequality, note that $r(\mathbf{a} + \mathbf{e}_i)/r(\mathbf{a}) \in [1, i]$, since if $\mathbf{a}$ is increased by $\mathbf{e}_i$, then the numerator of $r(\mathbf{a})$ is multiplied by i, whereas the denominator of $r(\mathbf{a})$ is multiplied by a divisor of i. In particular, $r(\cdot)$ is an increasing function. Finally,

$$r(\mathbf{a}) \le r(\mathbf{b} + \mathbf{e}_j) \le jr(\mathbf{b}) \le nr(\mathbf{b}) ,$$

completing the proof. ■

The probabilistic use of the last lemma is given by

Lemma 3. *Let* $\mathbf{C}_n = (C_1(n), \dots, C_n(n))$ *have the distribution* (7), *let* $\{Z_j, j \ge 1\}$ *be independent Poisson random variables with* $\mathbb{E}Z_j = \theta/j$, *and set* $\mathbf{Z}_n = (Z_1, \dots, Z_n)$. *Then there is a coupling for which for every* n

$$0 \le \log r(\mathbf{C}_n) \equiv \log P_n - \log O_n \le \log n + \log r(\mathbf{Z}_n) . \tag{28}$$

Proof. Use the result described in Theorem 1, which guarantees the existence of a coupling satisfying $\mathbf{C}_n \le \mathbf{Z}_n + \mathbf{e}_{J_n}$, where $1 \le J_n \le n$. Now apply Lemma 2. ■

The next lemma is a calculation for the Poisson process. The analogous result for uniformly distributed random permutations (that is, $\theta = 1$) was proved directly by DeLaurentis and Pittel [14].

Lemma 4. *As* $n \to \infty$,

$$\mathbb{E}\, log\, r(\mathbf{Z}_n) = O(\log n(\log\log n)^2)\,. \tag{29}$$

Proof. For $1 \le k \le n$, define a function d_{nk} by

$$d_{nk}(\mathbf{a}) = \sum_{j \le n; k|j} a_j\,,$$

and note that $D_{nk} \equiv d_{nk}(\mathbf{Z}_n)$ has a Poisson distribution satisfying

$$\mathbb{E}D_{nk} = \theta \sum_{j \le n; k|j} 1/j = O(\log n/k)\,, \tag{30}$$

and

$$\mathbb{E}D_{nk}(D_{nk} - 1) = \Big(\theta \sum_{j \le n; k|j} 1/j\Big)^2 = O(\log^2 n/k^2)\,, \tag{31}$$

uniformly in $k \le n$. Note that since $(D_{nk} - 1)^+ \le D_{nk}$, it follows from (30) that

$$\begin{aligned}\mathbb{E} \sum_{k \le \log n} \log k(D_{nk} - 1)^+ &\le c \log n \sum_{k \le \log n} \log k/k \\ &= O(\log n(\log\log n)^2)\,.\end{aligned}$$

Similarly, since $(D_{nk} - 1)^+ \le D_{nk}(D_{nk} - 1)/2$, it follows from (31) that

$$\begin{aligned}\mathbb{E} \sum_{k > \log n} \log k(D_{nk} - 1)^+ &\le c \log^2 n \sum_{k > \log n} \log k/k^2 \\ &= O\,(\log n \log\log n)\,.\end{aligned}$$

Combining these two estimates, we see that

$$\mathbb{E} \sum_{k \ge 1} \log k(D_{nk} - 1)^+ = O(\log n(\log\log n)^2)\,. \tag{32}$$

Finally

$$\begin{aligned}\mathbb{E} \log r(\mathbf{Z}_n) &= \mathbb{E} \sum_{p \text{ prime}} \sum_{s \ge 1} \log p(d_{np^s}(\mathbf{Z}_n) - 1)^+ \\ &\le \mathbb{E} \sum_{k \ge 1} \log k(D_{nk} - 1)^+\,,\end{aligned}$$

the right-hand side being $O(\log n(\log\log n)^2)$ by (32). This completes the proof. ∎

The generalized version of the Erdös–Turán Law for the Ewens sampling formula is

Theorem 8. *As* $n \to \infty$,

$$\frac{\log O_n - \frac{\theta}{2}\log^2 n}{\sqrt{\frac{\theta}{3}\log^3 n}} \Rightarrow N(0,1)\,. \tag{33}$$

Remark. There are several proofs of the $\theta = 1$ version of this result, among them Best [10], Kolchin [31, 32, p. 61], Bovey [13], DeLaurentis and Pittel [14], and Stein [38]. We are aware of, but have not seen, the proof of Pavlov [36].

Proof. First we combine (28) and (29) to conclude that

$$0 \le \frac{\mathbb{E}(\log P_n - \log O_n)}{\log^{3/2} n} = O\left(\frac{(\log\log n)^2}{\log^{1/2} n}\right),$$

from which it follows that the theorem will be proved if we establish that

$$\frac{\sum_{j=1}^{n} C_j(n)\log j - \frac{\theta}{2}\log^2 n}{\sqrt{\frac{\theta}{3}\log^3 n}} \Rightarrow N(0,1)\,. \tag{34}$$

As in the previous examples, we prove the result with (dependent) $C_j(n)$ replaced by (independent) Z_j, and show that the error in this approximation is negligible. Observe that $\sum_{j=1}^{n} \log j\,\mathbb{E}Z_j \sim \theta\log^2 n/2$ and $\sum_{j=1}^{n} \log^2 j/j \sim \log^3 n/3$. Direct methods (or an appeal to the Lindeberg–Feller conditions) then establish that

$$\frac{\sum_{j=1}^{n} Z_j\log j - \frac{\theta}{2}\log^2 n}{\sqrt{\frac{\theta}{3}\log^3 n}} \Rightarrow N(0,1)\,. \tag{35}$$

Using Lemma 1 again, we see that the absolute value of the error R_n in the approximation of the left side of (34) by the left side of (35) is

$$\begin{aligned}|R_n| &= \frac{\left|\sum_{j=1}^{n} \log j(C_j - Z_j)\right|}{\sqrt{\frac{\theta}{3}\log^3 n}}\\ &\le \sqrt{3}\,\frac{\sum_{j=1}^{n} |C_j - Z_j|}{\sqrt{\theta\log n}}\\ &\equiv \sqrt{3}R_n^* \to_P 0\,.\end{aligned}$$

This completes the proof. ∎

4. NONLINEAR FUNCTIONALS

The examples in Section 2 have studied the behavior of linear functionals of the cycle counting process. The Erdös–Turán law in Section 3 starts with least common multiple, a nonlinear functional, but is proved by comparison with a linear functional. Other nonlinear functionals are also of interest. Motivated by a result of Wilf [41] for uniform random permutations, we study the behavior of the number of different sizes of cycles in a random permutation. Theorem 9 gives a limit distribution with no rescaling, in contrast to the Theorems of Section 2, which involve rescaling by $\sqrt{\log n}$. We begin with a preliminary lemma that builds on the results of Theorem 1:

Lemma 5. *As* $n \to \infty$,

$$\mathbb{E}(Z_{J_n}) = O\left(\frac{\log n}{n}\right).$$

In fact,

$$\mathbb{E}(Z_{J_n}) \le \frac{3\theta}{n}(1 + \theta + \theta \log n) + \frac{\theta}{\theta + n}.$$

Proof. By conditioning on the event $\{J_n = j\}$ and using the definitions in (10), (11), and (14), we see that

$$\mathbb{E}(Z_{J_n} \mid J_n = j) \le \mathbb{E}\xi_{n-2j+1}(1 - \xi_{n-2j+2}) \cdots (1 - \xi_{n-j}) + \mathbb{E}\xi_{n+1} + \mathbb{E}Z_j, \quad (36)$$

where we define $\xi_m \equiv 0$ if $m < 1$.

From (7) (see also Watterson [40]) we have

$$\mathbb{E}C_j(n) = \frac{\theta}{j} \frac{n(n-1)\cdots(n-j+1)}{(\theta+n-j)\cdots(\theta+n-1)}. \quad (37)$$

From (14) and (37) it follows that

$$\mathbb{P}(J_n = j) = \frac{j\mathbb{E}C_j(n)}{n}, \quad (38)$$

so that from (19)

$$\begin{aligned}\sum_{j=1}^{n} \mathbb{E}Z_j\, \mathbb{P}(J_n = j) &= \sum_{j=1}^{n} \frac{\theta}{j} \frac{j\,\mathbb{E}C_j(n)}{n} \\ &= \frac{\theta}{n} \sum_{j=1}^{n} \mathbb{E}C_j(n) \\ &= \frac{\theta}{n} \mathbb{E}K_n \\ &\le \frac{\theta}{n}(1 + \theta + \theta \log n).\end{aligned}$$

Using (14), (19), and (37) once more, we find that

$$\begin{aligned}\sum_{j=1}^{\lfloor n/2\rfloor} \mathbb{E}\xi_{n-2j+1}(1-\xi_{n-2j+2})\cdots(1-\xi_{n-j})\,\mathbb{P}(J_n=j) &= \theta \sum_{j=1}^{\lfloor n/2\rfloor} \frac{2j}{n}\,\frac{1}{n-j}\,\mathbb{E}C_{2j}(n)\\ &\le \theta \sum_{j=1}^{\lfloor n/2\rfloor} \frac{1}{n-j}\,\mathbb{E}C_{2j}(n)\\ &\le \frac{2\theta}{n} \sum_{j=1}^{\lfloor n/2\rfloor} \mathbb{E}C_{2j}(n)\\ &\le \frac{2\theta}{n} \sum_{j-1}^{n} \mathbb{E}C_j(n)\\ &= \frac{2\theta}{n}\,\mathbb{E}K_n\\ &\le \frac{2\theta}{n}\,(1+\theta+\theta\log n)\,.\end{aligned}$$

Averaging the inequality (36) over the distribution of J_n, and using the two preceeding inequalities and the fact that $\mathbb{E}\xi_{n+1} = \theta/(\theta+n)$ completes the proof of the Lemma. ■

Our interest is in the quantity D_n, the difference between the number of cycles and the number of distinct cycle lengths in a random permutation. By definition, we have

$$D_n = \sum_{j\le n} (C_j(n) - 1(C_j(n)\ge 1)) = \sum_{j\le n} (C_j(n)-1)^+ \,.$$

Let $\{Z_j,\ j\ge 1\}$ be the Poisson random variables defined by the coupling in Theorem 1. Our analysis of D_n is based on the observation that for any $I \subseteq \{1,\ldots,n\}$,

$$\sum_{j\in I} (C_j(n)-1)^+ \le \sum_{j\in I} (Z_j-1)^+ + 1(Z_{J_n}\ge 1)\,. \tag{39}$$

Theorem 9. *As $n\to\infty$, we have*

$$D_n \Rightarrow D \equiv \sum_{j\ge 1} (Z_j-1)^+ \,, \tag{40}$$

and

$$\mathbb{E}D_n \to \mathbb{E}D \equiv \sum_{j\ge 1} \left(\frac{\theta}{j} - 1 + \exp(-\theta/j)\right). \tag{41}$$

Proof. Define $D'_n = \sum_{j\le n}(Z_j-1)^+$. Clearly, $D'_n \Rightarrow D$ as $n\to\infty$. To establish (40), it therefore suffices to show that $D_n - D'_n \to_P 0$. By Theorem 2, for any $1\le b\le n$ we can choose a coupling of $(C_1(n),\ldots,C_n(n))$ and $(Z_1, Z_2,\ldots)$ so that

$$\mathbb{P}\left(\left|\sum_{j\le b}((C_j(n)-1)^+-(Z_j-1)^+)\right|>0\right)\le d_b(n)\,.$$

Using (39) with $I=\{b+1,\dots,n\}$,

$$\begin{aligned}\left|\sum_{j=b+1}^{n}((C_j(n)-1)^+-(Z_j-1)^+)\right| &\le \sum_{j=b+1}^{n}(C_j(n)-1)^+ + \sum_{j=b+1}^{n}(Z_j-1)^+\\ &\le 2\sum_{j=b+1}^{n}(Z_j-1)^+ + 1(Z_{J_n}\ge 1)\,.\end{aligned} \tag{42}$$

Since the Z_j are Poisson distributed, it follows that

$$\begin{aligned}\mathbb{E}\sum_{j=b+1}^{n}(Z_j-1)^+ &\le \mathbb{E}\sum_{j=b+1}^{n}Z_j(Z_j-1)/2\\ &=\sum_{j=b+1}^{n}\theta^2/(2j^2)\\ &\le \theta^2/(2b)\,,\end{aligned}$$

whereas from Lemma 5, we have $\mathbb{E}(Z_{J_n})=O(\log n/n)$.

Finally, we may choose $b\equiv b(n)\to\infty$ in such a way that $b/n\to 0$; this will guarantee that $d_b(n)\to 0$, and also that the quantity on the right of (42) converges to 0 in probability. This establishes (40).

To obtain (41) from (40), we show that the sequence $\{D_n,\ n\ge 1\}$ is uniformly integrable. Applying (39) with $I=\{1,\dots,n\}$, we see that

$$\begin{aligned}0\le D_n &\le \sum_{j=1}^{n}(Z_j-1)^+ + 1(Z_{J_n}\ge 1)\\ &\le \sum_{j\ge 1}(Z_j-1)^+ + 1\,.\end{aligned} \tag{43}$$

Since the random variable on the right side of (43) has finite mean, we conclude that indeed $\{D_n,\ n\ge 1\}$ is uniformly integrable. ■

Remark. Wilf [41] proved (41) in the special case $\theta=1$ by analytical methods.

Remark. It follows immediately from Theorems 3 and 9 that the number of distinct cycle lengths in a random permutation has asymptotically a Normal distribution with mean and variance $\theta\log n$.

5. RANDOM MAPPINGS

In this section, we study the collection of n^n mappings of the set $\{1,\dots,n\}$ to itself, under the assumption that all such mappings are equally likely. Each mapping partitions the set $\{1,\dots,n\}$ into components (integers l and m being in the same component if some iterate of l is equal to some iterate of m). In particular, we study the behavior of the numbers $C_1(n)$, $C_2(n), \ldots$ of components of sizes $1, 2, \ldots$. The results in Theorems 3, 4, and 5, specialized to $\theta = 1/2$, also hold in the random mapping setting. The theorem corresponding to Theorem 3 is the central limit theorem for the number of components of a random mapping, first proved by Stepanov [39]. The theorem corresponding to Theorem 5 is the functional central limit theorem for random mappings, due originally to Hansen [26]. In the random mapping versions of Theorems 6 and 7 we also need to replace the Poisson process $\{Z_j, j \geq 1)$, $\mathbb{E}Z_j = \theta/j$ with a Poisson process $\{Z_j, j \geq 1\}$ in which $\mathbb{E}Z_j = e^{-j}/j \sum_{i=0}^{j-1} j^i/i!$. Related results for linear combinations of component sizes and the smallest component sizes are addressed by Kolchin [32, pp. 85 ff].

The crucial ingredients for the proofs of these results are given in the following section. In the case of permutations, the result $R_n^* \rightarrow_P 0$ followed easily from the coupling given in Theorem 1. For random mappings, the result $R_n^* \rightarrow_P 0$ uses instead a combination of total variation approximations, given by Theorem 10, and moment estimates from Lemma 6.

A. The Components of a Random Mapping

Harris [28] showed that the probability that a random mapping has component index $(a_1, a_2, \ldots, a_n)$ (that is, has a_j components of size j) is

$$\mathbb{P}(C_j(n) = a_j,\ j = 1, \ldots, n) = \frac{n!e^n}{n^n} \prod_{j=1}^{n} \frac{\lambda_j^{a_j}}{a_j!}\, 1\left\{\sum_{j=1}^{n} ja_j = n\right\}, \tag{44}$$

where

$$\lambda_j = \frac{e^{-j}}{j} \sum_{i=0}^{j-1} \frac{j^i}{i!}\,. \tag{45}$$

If follows immediately from (44) that

$$\mathbb{E}C_j(n) = \frac{n!e^n}{n^n} \frac{(n-j)^{n-j}}{e^{n-j}(n-j)!}\, \lambda_j\,. \tag{46}$$

Under the distribution (44), Kolchin [30] established that

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

where the Z_i are independent Poisson random variables with mean

$$\mathbb{E}Z_i = \lambda_i\,. \tag{47}$$

Arratia and Tavaré [4] established the following analog of Theorem 2:

Theorem 10. *Let* $(C_1(n), C_2(n), \ldots)$ *be the* component *counting process for random mappings, and let* $(Z_1, Z_2, \ldots)$ *be the Poisson process on* $\mathbb{N}$ *determined by*

(47). *For* $1 \leq b \leq n$ *let* $d_b(n)$ *be the total variation distance between* $(C_1(n), \ldots, C_b(n))$ *and* $(Z_1, \ldots, Z_b)$ *defined in* (4). *Then*

$$d_b(n) \to 0 \text{ if, and only if, } b = o(n) \tag{48}$$

Lemma 1 was crucial in our analysis of random permutations. The following two lemmas play this role in the context of random mappings.

Lemma 6. *Let* $Z_1, Z_2, \ldots$ *be independent Poisson random variables with* $\mathbb{E}(Z_i) = \lambda_i$, *and let* $(C_1(n), C_2(n), \ldots)$ *be the component counting process for a random mapping. Then for* $1 \leq b \leq n$ *and* $f \equiv n/b$,

$$\mathbb{E} \sum_{j=b+1}^{n} Z_j = O(\log f)\,, \tag{49}$$

and

$$\mathbb{E} \sum_{j=b+1}^{n} C_j(n) = O(\log f)\,. \tag{50}$$

Proof. We may write $\lambda_j = j^{-1}\,\mathbb{P}(\mathrm{Po}(j) < j)$, where $\mathrm{Po}(j)$ is a Poisson random variable with mean j. It follows from the central limit theorem that $\lambda_j \sim 1/(2j)$ as $j \to \infty$. Further (cf. Donnelly, Ewens, and Padmadisastra [16])

$$\sum_{j=1}^{\infty} \left(\frac{1}{2j} - \lambda_j\right) = \frac{1}{2}\log 2\,. \tag{51}$$

The inequality in (49) follows immediately, since

$$\sum_{j=b+1}^{n} \lambda_j = \sum_{j=b+1}^{n} \frac{1}{2j} + \sum_{j=b+1}^{n} \left(\lambda_j - \frac{1}{2j}\right) = O(\log f) + O(1)\,,$$

using (51) and the fact that $\sum_{j=b+1}^{n} 1/j \leq \log(n/b) \equiv \log f$.

To establish (50), note first that from (46)

$$\begin{aligned}\sum_{j=b+1}^{n} \mathbb{E}C_j(n) &= \frac{n!e^n}{n^n} \sum_{j=b+1}^{n} \frac{(n-j)^{n-j}}{e^{n-j}(n-j)!}\,\lambda_j \\ &\leq \frac{n!e^n}{n^n} \sum_{j=b+1}^{n} \frac{(n-j)^{n-j}}{e^{n-j}(n-j)!}\,\frac{1}{j} \\ &= \frac{n!e^n}{n^n} \sum_{r=0}^{n-b-1} \frac{r^r}{e^r r!}\,\frac{1}{n-r}\,.\end{aligned}$$

We will require $b/n \to 0$, so we may assume that n is sufficiently large that $(b+1)/n \leq 2/3$. The dominant contribution to the right-hand side of the previous inequality comes from the terms with $r \geq r_0 \equiv \lceil n/3 \rceil$, as may be seen by applying Stirling's formula to get

$$
\begin{aligned}
\sum_{r=r_0}^{n-b-1} \frac{r^r}{e^r r!} \frac{1}{n-r} &\le \frac{1}{\sqrt{2\pi}} \sum_{r=r_0}^{n-b-1} \frac{1}{(n-r)\sqrt{r}} \\
&\le \frac{1}{\sqrt{2\pi n}} \int_{r_0/n}^{1-b/n} \frac{dy}{(1-y)\sqrt{y}} \\
&\le \frac{1}{\sqrt{2\pi n}} \int_{1/3}^{1-b/n} \frac{dy}{(1-y)\sqrt{y}} \\
&\le \frac{\sqrt{3}}{\sqrt{2\pi n}} \int_{0}^{1-bn} \frac{dy}{1-y} \\
&= \frac{\sqrt{3}}{\sqrt{2\pi n}} \log(n/b) .
\end{aligned}
$$

A similar analysis shows that

$$\sum_{r=1}^{r_0-1} \frac{1}{(n-r)\sqrt{r}} \le \frac{3}{\sqrt{2\pi n}} .$$

Combining these estimates together with the term $r = 0$ shows that

$$\sum_{r=0}^{n-b-1} \frac{r^r}{e^r r!} \frac{1}{n-r} = O\left(\frac{1}{\sqrt{n}} \log(n/b)\right),$$

and (50) follows by another application of Stirling's Formula. ∎

The analog of Lemma 1 in the context of random mappings is contained in

Lemma 7. *There is a coupling of* $\{C_j(n), j \ge 1, n \ge 1\}$ *and* $\{Z_j, j \ge 1\}$ *such that*

$$R_n^* = \frac{\sum_{j=1}^{n} |C_j(n) - Z_j|}{\sqrt{\log n}} . \tag{52}$$

converges in probability to 0 *as* $n \to \infty$.

Proof. Observe that for any $1 \le b \le n$, we have

$$|R_n^*| \le R^*(b, n) \equiv \frac{\sum_{j=1}^{b} |C_j(n) - Z_j|}{\sqrt{\theta \log n}} + \frac{\sum_{j=b+1}^{n} Z_j}{\sqrt{\theta \log n}} + \frac{\sum_{j=b+1}^{n} C_j(n)}{\sqrt{\theta \log n}} \tag{53}$$

We write $R^*(b, n)$ in the form

$$R^*(b, n) = R_1 + R_2 + R_3 ,$$

and show that we may choose $b \equiv b(n)$ so that each term tends to 0 in probability as $n \to \infty$.

Choose a maximal coupling of $\mathbf{C}_b(n) \equiv (C_1(n), \ldots, C_b(n))$ and $\mathbf{Z}_b \equiv$

$(Z_1, \ldots, Z_b)$ and apply (6): for any $\epsilon > 0$,

$$\mathbb{P}(R_1 > \epsilon) \le \mathbb{P}(R_1 \neq 0) \le \mathbb{P}(\mathbf{C}_b \neq \mathbf{Z}_b) = d_b(n)\,. \tag{54}$$

Now extend the maximal coupling of $\mathbf{C}_b$ and $\mathbf{Z}_b$ to a coupling of $(C_1, \ldots, C_n)$ with $(Z_1, \ldots, Z_n)$. Markov's inequality together with the estimates in (49) and (50) show that if $f \equiv n/b$

$$\mathbb{P}(R_2 > \epsilon) \le \frac{c_1 \log f}{\epsilon\sqrt{\log n}}\,, \tag{55}$$

and

$$\mathbb{P}(R_3 > \epsilon) \le \frac{c_2 \log f}{\epsilon\sqrt{\log n}}\,. \tag{56}$$

Finally, we choose b to ensure that $d_b(n) \to 0$ and that the right side of (55) and (56) $\to 0$. We need

$$f \equiv n/b \to \infty;\ \frac{\log f}{\sqrt{\log n}} \to 0\,. \tag{57}$$

With such a choice of b, we see from Theorem 10 and Equations (54), (55), and (56) that $R^*(b, n) \to_P 0$ as $n \to \infty$, completing the proof. ∎

Now using Lemma 7 in place of Lemma 1, it is straightforward to modify the proofs of Theorems 3 through 7 to give proofs of the corresponding results for random mappings. We omit further details.

6. EXPLOITING UNIFORM BOUNDS

The previous sections have made no use of explicit bounds on total variation distances between the cycle or component counting processes and their respective Poisson limits; all that was exploited was the fact that $d_b(n) \to 0$ if $b/n \to 0$. In this section we use the more detailed information given by upper bounds on $d_b(n)$. The following result of Arratia, Barbour, and Tavaré [6, Theorem 5] complements Theorem 2:

Lemma 8. *Let $(C_1(n), C_2(n), \ldots)$ be the cycle counting process for the Ewens sampling formula with $\theta \ge 1$, and let $(Z_1, Z_2, \ldots)$ be the Poisson process on $\mathbb{N}$ determined by* (8). *For $1 \le b \le n$ let $d_b(n)$ be the total variation distance between $(C_1(n), \ldots, C_b(n))$ and $(Z_1, \ldots, Z_b)$. Then*

$$d_b(n) \le \frac{b\theta(\theta+1)}{\theta + n}\,. \tag{58}$$

A bound for $d_b(n)$ in the case $\theta < 1$ may be found in Arratia, Barbour, and Tavaré [7]. For the case $\theta = 1$, the result $d_b(n) \le 2b/n$ was proved by Diaconis and Pitman [15] and independently by Barbour [8]. When $\theta = 1$, the bound in (58) may be much improved. Arratia and Tavaré [3] show that if $b/n \to 0$, then $d_b(n) \to 0$ super-exponentially fast relative to n/b. In the next section we use the bounds in (8) to provide new results about the *cycle* structure of a random mapping.

A. The Cycles of a Random Mapping

Here we study some aspects of the structure of the cycles of a random mapping of $\{1, \ldots, n\}$ to itself. The core of a random mapping is the set of elements that are in cycles. In particular, the number N_n of elements in the core has distribution given by

$$\mathbb{P}(N_n = r) = \frac{r}{n} \prod_{l=1}^{r-1} \left(1 - \frac{l}{n}\right), r = 1, \ldots, n . \tag{59}$$

It follows directly from (59) that $N_n/\sqrt{n}$ converges in distribution to a random variable with density function $xe^{-x^2/2}$, $x > 0$. We let $C_j^* \equiv C_j^*(n)$ be the number of cycles of size j in the core of a random mapping, and let $C_j(r)$ be the number of cycles of size j in a *uniform* random permutation of r objects. The law $\mathscr{L}(C_j^*)$ of C_j^* is given by

$$\mathscr{L}(C_j^*(n)) = \sum_{r=1}^{n} \mathbb{P}(N_n = r)\mathscr{L}(C_j(r)) , \tag{60}$$

since, conditional on $N_n = r$ the random mapping restricted to its core is a uniformly distributed permutation on those r elements. It follows that

$$\mathbb{E}C_j^*(n) = \frac{1}{j} \prod_{l=1}^{j-1} \left(1 - \frac{l}{n}\right) . \tag{61}$$

For fixed j, $\mathbb{E}C_j^*(n) \to 1/j$, and

$$C_j^*(n) \Rightarrow Z_j , \tag{62}$$

where Z_j are independent Poisson random variables with mean $\mathbb{E}Z_j = 1/j$. Results (59) through (62) are classical; see, for example, Bollobás [12, p. 366].

Define $\mathbf{C}_b^*(n) = (C_1^*(n), \ldots, C_b^*(n))$, $\mathbf{Z}_b = (Z_1, \ldots, Z_b)$, and let $d_b^*(n)$ be the total variation distance between $\mathbf{C}_b^*(n)$ and $\mathbf{Z}_b$. Further, let $d_b(r)$ be the total variation distance between $\mathbf{Z}_b$ and the cycle counting process $\mathbf{C}_b(r) \equiv (C_1(r), \ldots, C_b(r))$ for uniform random permutations of r objects.

Theorem 11. *Let $\{Z_j, j \geq 1\}$ be independent Poisson random variables with means $\mathbb{E}Z_j = 1/j$. Then*

$$d_b^*(n) \to 0 \text{ if, and only if, } b = o(\sqrt{n}) . \tag{63}$$

Proof. The joint law version of (60) is

$$\mathscr{L}(\mathbf{C}_b^*(n)) = \sum_{r=1}^{n} \mathbb{P}(N_n = r)\mathscr{L}(\mathbf{C}_b(r)) ,$$

which implies that

$$d_b^*(n) \le \sum_{r=1}^{n} \mathbb{P}(N_n = r) d_b(r) . \tag{64}$$

The estimate (58) with $\theta = 1$ shows that $d_b(r) \le 2b/r$ so that

$$\begin{aligned} d_b^*(n) &\le \sum_{r=1}^{n} \frac{2b}{r} \cdot \frac{r}{n} \prod_{l=1}^{r-1} \left(1 - \frac{l}{n}\right) \\ &= \frac{2b}{n} \sum_{r=1}^{n} \prod_{l=1}^{r-1} \left(1 - \frac{l}{n}\right) \\ &\sim \frac{2b}{n} \sqrt{\frac{n\pi}{2}} , \end{aligned} \tag{65}$$

the last estimate coming from Bollobás [12, p. 114]. From (65), we see that $d_b^*(n) \to 0$ if $b = o(\sqrt{n})$.

To prove the converse, note first that for any $1 \le L \le M \le b$,

$$\|\mathscr{L}(\mathbf{C}_b^*(n)) - \mathscr{L}(\mathbf{Z}_b)\| \ge \|\mathscr{L}(C_L^* + \cdots + C_M^*) - \mathscr{L}(Z_L + \cdots + Z_M)\| .$$

Assume that $b/\sqrt{n} \ge \delta > 0$ for all n, and define $M = \lfloor \delta\sqrt{n} \rfloor$, $L = \lfloor \delta\sqrt{n}/2 \rfloor$. Define $X_n = Z_L + \cdots + Z_M$, $Y_n = C_L^* + \cdots + C_M^*$. X_n is a Poisson random variable satisfying

$$\mathbb{E}X_n = \sum_{j=L}^{M} 1/j \to \log 2 ,$$

so that $X_n \Rightarrow X$ which is a Poisson random variable with finite mean. Since

$$\mathbb{E}C_j^* \sim \frac{1}{j} e^{-j^2/(2n)}$$

for $L \le j \le M$, we see that

$$\mathbb{E}X - \mathbb{E}Y_n \to \epsilon \equiv \log 2 - \int_{\delta/2}^{\delta} \frac{1}{x} e^{-x^2/2} dx > 0 .$$

Now choose t such that $\mathbb{E}\min(X, t) > \mathbb{E}X - \epsilon/4$. Since

$$|\mathbb{E}\min(X, t) - \mathbb{E}\min(Y_n, t)| \le t\|\mathscr{L}(X) - \mathscr{L}(Y_n\|$$

it follows that

$$\begin{aligned} \liminf_{n\to\infty} \|\mathscr{L}(X_n) - \mathscr{L}(Y_n)\| &= \liminf_{n\to\infty} \|\mathscr{L}(X) - \mathscr{L}(Y_n))\| \\ &\ge \epsilon/(4t) > 0 , \end{aligned}$$

completing the proof. ■

Remark. The above proof of the necessity of the condition $b = o(\sqrt{n})$ relies on the discrepancy in the first moment of $C_j^*(n)$ and Z_j. Even if the Z_j were replaced by $Z_j(n)$, defined to be Poisson with $\mathbb{E}Z_j(n) = \mathbb{E}C_j(n)$, the condition $b = o(\sqrt{n})$ would still be necessary for $d_b^*(n) \to 0$. To see this, observe that $X_n \Rightarrow X$, $\mathbb{E}X - \mathbb{E}Y_n \to 0$ and $\mathbb{E}X^2 - \mathbb{E}Y_n^2 \to \epsilon^* > 0$. We thank Andrew Barbour for pointing this out.

Since the total number of components of a random mapping equals the total number of cycles, this result may be used to give another proof of Stepanov's [39] central limit theorem for the number of components of a random mapping. Theorems 3 through 9 of the previous sections have analogs for the core. One of these, the analog of Theorem 7, is tricky to state, so we give it explicitly as:

Theorem 12. *The joint distribution of the m smallest cycles of a random mapping may be approximated in terms of $\{Z_j,\ j \geq 1\}$, the Poisson process with means given by $\mathbb{E}Z_j = 1/j$, in the sense that the total variation distance d_m^* defined in* (25) *tends to* 0 *if, and only if, $\omega_n \equiv (\frac{1}{2}\log n - m)/\sqrt{\log n}$ satisfies $\omega_n \to \infty$.*

Proof. Similar to the proof of Theorem 7, but now to get $d_b(n) \to 0$ we need $b = 0(\sqrt{n})$, which explains the factor $\frac{1}{2}$ in the condition $m = \frac{1}{2}\log n - \omega_n\sqrt{\log n}$. ■

REFERENCES

[1] D. J. Aldous and J. W. Pitman, Brownian bridge asymptotics for random mappings. In preparation.

[2] R. Arratia, L. Goldstein, and L. Gordon, Poisson approximation and the Chen-Stein method. *Stat. Sci.*, **5**, 403–423 (1990).

[3] R. Arratia and S. Tavaré, The cycle structure of random permutations. *Ann. Probab.*, in press (1992a).

[4] R. Arratia and S. Tavaré, Functional discrete limit theorems for random mappings. In preparation (1992b).

[5] R. Arratia and S. Tavaré, Independent process approximations for random combinatorial structures. In preparation (1992c).

[6] R. Arratia, A. D. Barbour, and S. Tavaré, Poisson process approximations for the Ewens Sampling Formula. *Ann. Appl. Probab.*, in press (1992a).

[7] R. Arratia, A. D. Barbour, and S. Tavaré, On random polynomials over finite fields. In preparation (1992b).

[8] A. D. Barbour, Comment on a paper of Arratia, Goldstein and Gordon. *Stat. Sci.* **5**, 425–427 (1990).

[9] A. D. Barbour and P. G. Hall, On the rate of Poisson convergence. *Math. Proc. Cambridge Philos. Soc.* **95**, 473–480 (1984).

[10] M. R. Best, The distribution of some variables on a symmetric group. *Ned. Akad. Wetensch. Indag. Math. Proc. Ser. A*, **73**, 385–402 (1970).

[11] P. Billingsley, *Convergence of Probability Measures*, Wiley, New York, 1968.

[12] B. Bollobás, *Random Graphs*. Academic Press, New York, 1985.

[13] J. D. Bovey, An approximate probability distribution for the order of elements of the symmetric group. *Bull. London Math. Soc.*, **12**, 41–46 (1980).

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

[15] P. Diaconis and J. W. Pitman, Unpublished lecture notes, Statistics Department, University of Califorina, Berkeley, 1986.

[16] P. Donnelly, W. J. Ewens, and S. Padmadisastra, Random functions: exact and asymptotic results. *Adv. Appl. Probab.* **23**, 437–455 (1991).

[17] P. Donnelly, T. G. Kurtz, and S. Tavaré, On the functional central limit theorem for the Ewens Sampling Formula. *Ann. Appl. Probab.*, **1**, 539–545 (1991).

[18] P. Erdös and P. Turán, On some problems of statistical group theory. I. *Z. Wahrscheinlichkeitstheorie* **4**, 175–186 (1965).

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

[20] S. N. Ethier and T. G. Kurtz, *Markov Processes: Characterization and Convergence*, Wiley, New York, 1986.

[21] W. J. Ewens, The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.*, **3**, 87–112 (1972).

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

[23] P. Flajolet and A. M. Odlyzko, Random mapping statistics, in *Proc. Eurocrypt '89*, J.-J. Quisquater, Ed. *Lecture Notes* in *Computer Science*, **434**, Springer-Verlag, New York, 1990, pp. 329–354.

[24] P. Flajolet and M. Soria, Gaussian limiting distributions for the number of components in combinatorial structures. *J. Combinat. Th. Ser. A*, **53**, 165–182 (1990).

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

[26] J. C. Hansen, A functional central limit theorem for random mappings. *Ann. Probab.* **17**, 317–332 (1989).

[27] J. C. Hansen, A functional central limit theorem for the Ewens Sampling Formula. *J. Appl. Probab.* **27**, 28–43 (1990).

[28] B. Harris, Probability distributions related to random mappings. *Ann. Math. Stat.*, **31**, 1045–1062 (1960).

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

[30] V. F. Kolchin, A problem of the allocation of particles in cells and random mappings. *Theory Probab. Its Appl.*, **21**, 48–63 (1976).

[31] V. F. Kolchin A new proof of asymptotic lognormality of the order of a random substitution. *Proceedings Combinatorial and Asymptotical Analysis*, Krasnoyarsk State University Press, 1977, pp. 82–93. (In Russian)

[32] V. F. Kolchin, *Random Mappings*, Optimization Software, Inc., New York, 1986.

[33] R. Lidl and H. Niederreiter, *Introduction to Finite Fields and their Applications*, Cambridge University Press, Cambridge, England, 1986.

[34] N. Metropolis and G.-C. Rota, Witt vectors and the algebra of necklaces. *Adv. Math.*, **50**, 95–125 (1983).

[35] N. Metropolis and G.-C. Rota, The cyclotomic identity. *Contemp. Math.*, **34**, 19–27 (1984).

[36] A. I. Pavlov, On a theorem by Erdös and Turán. *Probl. Cybern.*, **64**, 57–66 (1980). (In Russian)

[37] L. A. Shepp and S. P. Lloyd, Ordered cycle lengths in a random permutation. *Trans. Am. Math. Soc.*, **121**, 340–357 (1966).

[38] C. Stein, The order of a random permutation. Unpublished manuscript.

[39] V. E. Stepanov, Limit distributions for certain charactrstics of random mappings. *Theory Probab. Its Appl.*, **14**, 612–626 (1969).

[40] G. A. Watterson, Models for the logarithmic species abundance distributions. *Theor. Popul. Biol.*, **6**, 217–250 (1974).

[41] H. Wilf, Three problems in combinatorial asymptotics. *J. Combinat. Th. Ser. A*, **35**, 199–207 (1983).

[42] P. Diaconis, M. McGrath and J. W. Pitman, Cycles and descents of random permutations, preprint (1992).

[43] A. D. Barbour, L. Holst and S. Janson, *Poisson Approximation*, Oxford University Press, Oxford, England, 1992.

Received July 22, 1991