Statistical Science
2016, Vol. 31, No. 1, 27–29
DOI: 10.1214/15-STS537
Main article DOI: 10.1214/15-STS529

Exploiting the Feller Coupling for the Ewens Sampling Formula

Richard Arratia, A. D. Barbour and Simon Tavaré

We congratulate Harry Crane on a masterful survey, showing the universal character of the Ewens sampling formula.

There are two grand ways to get a simple handle on the Ewens sampling formula; one is the Chinese restaurant coupling, and the other is the Feller coupling. Since Crane has discussed the Chinese Restaurant process, but not the Feller coupling, we will give a brief survey of the latter.

The Ewens sampling formula, given in Crane's (1), has an interpretation in terms of the cycle type of a random permutation of n objects. For $\theta = 1$, it is just Cauchy's formula, expressed in terms of the *fraction* of permutations of n objects that have exactly m_i cycles of order i, $1 \le i \le n$. For general θ, the power

$$\theta^{m_1+m_2+\cdots+m_n} = \theta^K$$

appearing in the formula, where K denotes the number of cycles, biases the uniform random choice of a permutation by weighting with the factor θ^K, the remaining factors involving θ merely reflecting the new normalization constant required to specify a probability distribution. We use the notation $(C_1(n), \dots, C_n(n))$ to denote a random object distributed according to the Ewens sampling formula, suppressing the parameter θ but making explicit the parameter n, so that, with Crane's notation (1),

$$\begin{aligned} &\mathbb{P}(C_1(n) = m_1, \dots, C_n(n) = m_n) \\ &\quad = p(m_1, \dots, m_n; \theta). \end{aligned} \tag{1}$$

Richard Arratia is Professor, Department of Mathematics, University of Southern California, 3620 S. Vermont Ave, KAP 104, Los Angeles, California 90089-2532, USA (e-mail: rarratia@usc.edu). A. D. Barbour is Professor Emeritus, Institut für Mathematik, Universität Zürich, Winterthurerstrasse 190, 8057 Zürich, Switzerland (e-mail: a.d.barbour@math.uzh.ch). Simon Tavaré is Professor, Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Centre for Mathematical Sciences, Wilberforce Road, Cambridge CB3 0WA, United Kingdom (e-mail: st321@cam.ac.uk).

The Feller coupling, motivated by the example in Feller ([6], page 815) is defined as follows. Take independent Bernoulli random variables ξ_i, $i = 1, 2, 3, \dots$, with the simple odds ratios $\mathbb{P}(\xi_i = 0)/\mathbb{P}(\xi_i = 1) = (i-1)/\theta$. Thus, $\mathbb{E}\xi_i = \mathbb{P}(\xi_i = 1) = \theta/(\theta + i - 1)$, and $\mathbb{P}(\xi_i = 0) = (i-1)/(\theta + i - 1)$. Say that an ℓ-spacing occurs in a sequence $a_1, a_2, \dots$, of zeros and ones, starting at position $i - \ell$ and ending at position i, if $a_{i-\ell}a_{i-\ell+1} \cdots a_{i-1}a_i = 10^{\ell-1}1$, a one followed by $\ell - 1$ zeros followed by another one. Then if, for each $\ell \ge 1$, we define

$$\begin{aligned} C_\ell(n) := {} & \text{the number of } \ell\text{-spacings in} \\ & \xi_1, \xi_2, \dots, \xi_{n-1}, \xi_n, 1, 0, 0, \dots, \end{aligned}$$

the joint distribution of $C_1(n), \dots, C_n(n)$ *is* the Ewens sampling formula, as per Crane's (1) and our (1). This can be seen directly, for the case $\theta = 1$: consider a random permutation of 1 to n, write the canonical cycle notation one symbol at a time, and let ξ_i indicate the decision to complete a cycle, when there is an i-way choice of which element to assign next. The general case $\theta > 0$ follows by biasing, with respect to θ^K: since $K = \xi_1 + \cdots + \xi_n$, and the $\xi_1, \dots, \xi_n$ are independent, biasing their joint distribution by $\theta^{\xi_1+\cdots+\xi_n} = \theta^{\xi_1} \cdots \theta^{\xi_n}$ preserves their independence and Bernoulli distributions, while changing the odds $\mathbb{P}(\xi_i = 0)/\mathbb{P}(\xi_i = 1)$ from $(i-1)/1$ to $(i-1)/\theta$.

Now, the wonderful thing that happens is that, with Y_ℓ *defined* to be the number of ℓ-spacings in the infinite sequence $\xi_1, \xi_2, \dots$, it turns out that $Y_1, Y_2, \dots$ are mutually independent, and that Y_ℓ is Poisson distributed, with $\mathbb{E}Y_\ell = \theta/\ell$, as in formula (11) in Section 3.8. This shows that the Ewens sampling formula is closely related to the simpler independent process $Y_1, Y_2, \dots, Y_n$. Explicitly, let R_n be the position of the rightmost one in $\xi_1, \xi_2, \dots, \xi_{n-1}, \xi_n$—noting that always $\xi_1 = 1$ so R_n is well-defined—and let $J_n := (n+1) - R_n$. We have

$$C_\ell(n) \le Y_\ell + 1(J_n = \ell), \qquad 1 \le \ell \le n, \tag{2}$$

with contributions to strict inequality whenever, for some $1 \le \ell \le n$, an ℓ-spacing occurred in $\xi_1, \xi_2, \dots$ starting at $i - \ell$ and ending at $i > n$.