Statistical Science
2016, Vol. 31, No. 1, 37–39
DOI: 10.1214/15-STS544
Main article DOI: 10.1214/15-STS529
© Institute of Mathematical Statistics, 2016

Rejoinder: The Ubiquitous Ewens Sampling Formula

Harry Crane

The main article and extended discussion point to Ewens's sampling formula (ESF) as one of a few essential probability distributions. Arratia, Barbour and Tavaré explain the emergence of ESF by the Feller coupling and also touch on number theoretic considerations; Feng provides deeper background on diffusion processes and nonequilibrium versions of ESF; and McCullagh regales us with a story from the works of Fisher and Good, putting historical context around the more specialized topics covered by Favaro and James and Teh. The breadth of these comments exemplifies the expansive sphere of influence of Ewens's sampling formula on integer partitions, Ewens's distribution on set partitions, and the Ewens process. I thank all of the discussants for their participation in this important survey.

For the most part, these contributions bolster my main thesis which, in the words of Arratia, Barbour and Tavaré, emphasizes the *universal character of the Ewens sampling formula*. As McCullagh notes, the contents and subsequent discussion comprise an *impressive list stretching from literary studies to population genetics and probabilistic number theory*. Both comments accord with my opening remark that *Ewens's sampling formula exemplifies the harmony of mathematical theory, statistical application, and scientific discovery*. As a whole, however, the discussion skews disproportionately toward Bayesian nonparametrics in a way that works against the theme of *ubiquity*. I attempt to rebalance the conversation in these final pages.

1. EWENS'S SAMPLING FORMULA IN MODERN STATISTICS

Wherever random partitions appear, with few exceptions, so does Ewens's sampling formula. Teh compares its *inevitability* to that of the Gaussian distribution for real-valued sequences, and McCullagh makes a further analogy between the Ewens process and the Poisson process for events in time or space. Its tangible connections to population genetics, inductive inference, stochastic process theory, prime factorization, and statistical applications earn Ewens's sampling formula and the Poisson–Dirichlet distribution a place alongside the Bernoulli, Gaussian, and Poisson in the pantheon of probability distributions.

The applicability of Ewens's sampling formula is neither limited to specific methods nor tied to ongoing trends: Teh centers his commentary around contemporary topics in machine learning and big data, Favaro and James deal with problems in survival modeling and species sampling, and McCullagh showcases the adaptability of ESF with an enlightening application to a problem considered by Fisher three decades before Ewens's discovery. As McCullagh details, Ewens's sampling formula and its derivatives, the Ewens distribution and Ewens process, could have—indeed, should have—been first discovered in a purely parametric context, when data sets were small and computers were in their infancy.

McCullagh rightly identifies Ewens's process as *one of a small number of processes that deserves to be a central part of the statistical curriculum*. Indeed, there are compelling reasons to teach ESF at every level of statistics, and yet it is often reserved for special topics or not covered at all. Its most salient features, namely, exchangeability, sampling consistency, and noninterference, highlight subtleties that do not arise in i.i.d. sampling models and which can be covered without any need to delve into population genetics, stochastic processes, or Bayesian nonparametrics.

2. EWENS'S SAMPLING FORMULA AND BAYESIAN NONPARAMETRICS

Of the three commentaries covering statistical elements of ESF, two (Favaro and James, Teh) focus on recent work in Bayesian nonparametrics while the other (McCullagh) presents an application from seventy years ago. Together these comments fit into a

Harry Crane is Assistant Professor of Statistics & Biostatistics, Rutgers, the State University of New Jersey, 110 Frelinghuysen Road, Room 501, Piscataway, New Jersey 08854, USA (e-mail: hcrane@stat.rutgers.edu).