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

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).

broader, but misleading, narrative that Bayesian nonparametrics is the lifeblood of ESF in present-day statistical research. Though several authors do build substantially on the prior work of Ewens, Kingman, and Pitman, for example, Ishwaran and James's [10] work on the generalized Chinese restauarant process, Favaro, et al.'s [8] analysis of conditional sampling formulas, and Ruggiero and Walker's [12] study of the Fleming–Viot process, the Dirichlet process prior remains the primary mechanism by which Ewens's sampling formula arises in Bayesian nonparametrics. I have two major comments regarding how this connection is covered in the larger literature.

First, of all the recent surveys cited by Teh ([8, 11, 12, 13, 14, 21, 22] in Teh's numbering), only one [14], page 108, acknowledges Ewens's 1972 article or refers to Ewens's sampling formula by name. This tendency isolates the occurrence of ESF in Bayesian nonparametrics from the rest of the literature, fostering the impression that ESF is a byproduct of purely nonparametric Bayesian concerns. Second, the Dirichlet process is primarily chosen to address practical concerns of *tractability [and] computational convenience* ([9], page 37), which sell short the ESF's more critical statistical and inferential properties (Sections 3 and 7). Both of these oversights undermine the significance of the Ewens family of distributions: the first completely ignores the larger body of work on ESF and the second presents ESF merely as a quick fix for computational challenges.

3. THREE VIGNETTES ON EWENS'S SAMPLING FORMULA IN POPULATION GENETICS

While it is true that Bayesian nonparametrics is *one of the most active areas of statistical research*, the field of population genetics provides the primary context and is the most prominent venue for ESF. Notwithstanding Feng's account, which provides an insightful overview of how variants of ESF arise by diffusion process approximations, the population genetics angle warrants much more attention than it has received so far. Below I touch on three direct consequences of ESF in population genetics.

First, Ewens's derivation had an immediate impact on mutation rate estimation. Before [7], geneticists estimated θ, or functions of θ, from the empirical allele frequencies. Ewens showed that the number K of alleles is a sufficient statistic for θ, indicating that these early procedures used precisely the wrong part of the data in estimation of θ. This is a rare, and perhaps unique, example of a case where a previously unsuspected sufficient statistic changed standard inference procedures.

Second, some geneticists, including Wright [13], claimed that in the selectively neutral case all alleles observed in a sample should have approximately equal frequencies. Ewens's sampling formula shows that this is the least likely outcome under selective neutrality. The two main reasons for this phenomenon are simple random sampling and history—older alleles have a greater probability of reaching a high frequency than alleles that have recently arisen by mutation. This observation is relevant when testing whether data from a sample of genes supports the neutrality hypothesis.

The third and most lasting effect of Ewens's sampling formula is that it partially influenced Kingman's development of the coalescent [11], now the main vehicle for research in population genetics. The coalescent leads not only to a beautiful mathematical theory, which still provides the most elegant derivation of ESF, but also to a practical scientific framework which has moved the field of theoretical population genetics toward largely retrospective questions like: "When did the most recent ancestor of all humans alive today live?" and "How can we detect the signatures of past selective events in contemporary genomes?"

4. OTHER INSTANCES OF EWENS'S SAMPLING FORMULA

4.1 Independent Process Approximation and the Feller Coupling

Arratia, Barbour, and Tavaré expound a clear and well-motivated account of how Ewens's sampling formula emerges from the Feller coupling, which rightly deserves a place in the main survey alongside the Chinese restaurant process (CRP). The Feller coupling is more mathematically natural than the CRP construction, and it also illustrates the powerful technique of approximating statistics of combinatorial structures using independent processes; see [2].

4.2 Markov Survival Processes

Favaro and James discuss a connection between Ewens's sampling formula and neutral to the right survival models in Bayesian nonparametrics. Dempsey and McCullagh [6] observe the same connection but without resorting to the Bayesian nonparametrics framework. In the so-called *pilgrim process*, risk sets evolve according to an asymmetric version of Aldous's beta-splitting model [1] with parameter $\beta > -1$. The

$\beta = 1$ case yields a random partition distributed according to Ewens's distribution which, upon extension to recurrent events, elicits a connection to the so-called Indian buffet process.

4.3 Scale-Free Interaction Networks

The family of Ewens distributions also comes up in ongoing work on statistical network analysis. The degree distributions of many observed networks behave according to a power law, that is, the proportion p_k of vertices with degree $k \geq 1$ grows like $p_k \sim k^{-\gamma}$ as $k \to \infty$ for some $\gamma > 1$. Barabási and Albert's [3] preferential attachment model is the most widely cited generating mechanism for power-law networks, but its dynamics do not translate to a viable statistical model for two important reasons. First, its dynamics are too rigid to adequately reflect how most networks form, and its lack of exchangeability often prevents inference beyond selected summary statistics. Second, the preferential attachment dynamics can only explain power-law behavior in the range $\gamma > 2$, but Crane and Dempsey [4] have recently found that many networks formed by repeated interactions within a population exhibit power-law behavior with exponent in the complementary range $1 < \gamma < 2$. Based on the Ewens–Pitman two-parameter family (Section 5.1), we have put forth a new model that produces a network with the power-law exponent in the correct range. Our model is a precursor to the broader framework of *edge exchangeable network models* [5], which is the correct notion of invariance for many network data sets.

REFERENCES

[1] ALDOUS, D. (1996). Probability distributions on cladograms. In *Random Discrete Structures* (*Minneapolis*, *MN*, 1993). *IMA Vol. Math. Appl.* **76** 1–18. Springer, New York. MR1395604

[2] ARRATIA, R. and TAVARÉ, S. (1994). Independent process approximations for random combinatorial structures. *Adv. Math.* **104** 90–154. MR1272071

[3] BARABÁSI, A.-L. and ALBERT, R. (1999). Emergence of scaling in random networks. *Science* **286** 509–512. MR2091634

[4] CRANE, H. and DEMPSEY, W. (2015). Atypical scaling behavior persists in real world interaction networks. Available at arXiv:1509.08184.

[5] CRANE, H. and DEMPSEY, W. (2015). Edge exchangeable network models and the power law. Unpublished manuscript.

[6] DEMPSEY, W. and MCCULLAGH, P. (2015). The pilgrim process. Available at arXiv:1412.1490.

[7] EWENS, W. J. (1972). The sampling theory of selectively neutral alleles. *Theoret. Population Biology* **3** 87–112. MR0325177

[8] FAVARO, S., LIJOI, A. and PRÜNSTER, I. (2013). Conditional formulae for Gibbs-type exchangeable random partitions. *Ann. Appl. Probab.* **23** 1721–1754. MR3114915

[9] GHOSAL, S. (2010). The Dirichlet process, related priors and posterior asymptotics. In *Bayesian Nonparametrics* (N. L. Hjort, C. Holmes, P. Müller and S. G. Walker, eds.) 35–79. Cambridge Univ. Press, Cambridge. MR2730660

[10] ISHWARAN, H. and JAMES, L. F. (2003). Generalized weighted Chinese restaurant processes for species sampling mixture models. *Statist. Sinica* **13** 1211–1235. MR2026070

[11] KINGMAN, J. F. C. (1982). The coalescent. *Stochastic Process. Appl.* **13** 235–248. MR0671034

[12] RUGGIERO, M. and WALKER, S. G. (2009). Bayesian nonparametric construction of the Fleming–Viot process with fertility selection. *Statist. Sinica* **19** 707–720. MR2514183

[13] WRIGHT, S. (1978). *Evolution and the Genetics of Populations* **4**. Univ. Chicago Press, Chicago, IL.