ENCYCLOPEDIA OF STATISTICAL SCIENCES

UPDATE VOLUME 2

S. KOTZ, C.B. READ, D.L. BANKS (eds.)

A WILEY-INTERSCIENCE PUBLICATION
John Wiley & Sons, Inc. 1998
NEW YORK • CHICHESTER • WEINHEIM • BRISBANE • SINGAPORE • TORONTO

[48] Therneau, T. M., Grambsch, P. M., and Fleming, T. R. (1990). Martingale-based residuals for survival models. *Biometrika*, **77**, 147–160.

[49] Trussell, J., Hankinson, R., and Tilton, J., eds. (1992). *Demographic Applications of Event History Analysis*. Clarendon Press, Oxford. (A collection of demographic papers including three "contests," where pairs of research teams address the same topics using the same data.)

[50] Trussell, J. and Richards, T. (1985). Correcting for unmeasured heterogeneity in hazard models using the Heckman–Singer procedure. In *Sociological Methodology 1985*, N. B. Tuma, ed. Jossey Bass, San Francisco, pp. 242–276.

[51] Tuma, N. B. and Hannan, M. T. (1984). *Social Dynamics. Models and Methods*. Academic Press, Orlando, Fla. (A pioneering book.)

[52] Tuma, N. B., Hannan, M. T., and Groeneveld, L. P. (1976). Dynamic analysis of event histories. *Amer. J. Sociol.*, **84**, 820–854. (Introduced the term "event history.")

[53] Vaupel, J. W., Manton, K. G., and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, **16**, 439–454. (The original paper on frailty models.)

[54] Voelkel, J. G. and Crowley, J. (1984). Nonparametric inference for a class of semi-Markov models with censored observations. *Ann. Statist.*, **12**, 142–160.

[55] Yamaguchi, K. (1991). *Event History Analysis*. Sage Publications, Newbury Park, Calif.

[56] Zeger, S. L. and Karim, R. M. (1991). Generalized linear models with random effects: a Gibbs sampler approach. *J. Amer. Statist. Ass.*, **86**, 79–86. (Bayesian inference in generalized linear models using Markov-chain Monte Carlo methods.)

[57] Zucker, D. M. and Karr, A. F. (1990). Nonparametric survival analysis with time-dependent covariate effects: a penalized partial likelihood approach. *Ann. Statist.*, **18**, 329–353.

(SOCIOLOGY, STATISTICS IN)

GERMÁN RODRÍGUEZ

EWENS SAMPLING FORMULA

The Ewens sampling formula (ESF) describes a specific probability for the partition of a positive integer into parts. The distribution contains one parameter, usually denoted by θ. For the case $\theta = 1$ the distribution is quite old, going back in effect to Cauchy [7], since it then describes the partition into cycles of a random permutation, with each possible permutation being equally likely. The distribution arises in a wider variety of combinatorial objects than permutations, and in many scientific disciplines.

DISTRIBUTION AND MOMENTS

The ESF is most easily described in terms of sequential sampling of animals from an infinite collection of distinguishable species (Fisher et al. [14], McCloskey [25], Engen [11]). We use this example throughout, except where other specific applications are discussed. Suppose that the species have random frequencies $\boldsymbol{P} = (P_1, P_2, \ldots)$ satisfying

$$0 < P_i < 1, \quad i = 1, 2, \ldots, \qquad \sum_{i=1}^{\infty} P_i = 1. \tag{1}$$

Conditional upon $\boldsymbol{P}$, the species types of the animals are assumed to be independent, any animal being of species i with probability P_i. In what follows, we assume that the random vector $\boldsymbol{P}$ is defined by

$$P_1 = W_1,$$
$$P_r = (1 - W_1)(1 - W_2)\cdots(1 - W_{r-1})W_r, \quad r = 2, 3, \ldots, \tag{2}$$

where the $W_1, W_2, \ldots$ are i.i.d. with density $\theta(1 - x)^{\theta - 1}$, $0 < x < 1$, $0 < \theta < \infty$. This implies a nonmultinomial distribution for the numbers of different species observed.

Suppose a sample of n animals is taken. We can describe the species partition of these animals in two ways. First, we can record the counts $C_j(n)$, the number of species represented by j animals. The vector $\boldsymbol{C}_n = (C_1(n), \ldots, C_n(n))$ satisfies $\sum_{j=1}^{n} jC_j(n) = n$, and if K_n is the number of distinct species to appear in the sample, $K_n = \sum_{j=1}^{n} C_j(n)$. Rather more information is obtained by recording $A_1(n)$, the number of animals of the first species to appear, $A_2(n)$, the number of animals of the second species to appear, and so on.

Under the assumption that the random vector $\boldsymbol{P}$ has distribution given by (2), the ESF (Ewens

[12]) gives the distribution of the vector $\boldsymbol{C}_n$ as

$$\mathbb{P}[\boldsymbol{C}_n = \boldsymbol{a}_n] = \frac{n!}{\theta_{(n)}} \prod_{j=1}^{n} \left(\frac{\theta}{j}\right)^{a_j} \frac{1}{a_j!}, \quad (3)$$

where $\theta_{(n)} = \theta(\theta + 1)\cdots(\theta + n - 1)$ and $\boldsymbol{a}_n = (a_1, a_2, \ldots, a_n)$ is a vector of nonnegative integers satisfying $a_1 + 2a_2 + \cdots + na_n = n$. Similarly, the distribution of K_n is

$$\mathbb{P}[K_n = k] = S(n,k)\theta^k/\theta_{(n)}, \quad (4)$$

where $S(n,k)$ is the coefficient of θ^k in $\theta_{(n)}$, that is, is the absolute value of a Stirling number* of the first kind (*see* STIRLING DISTRIBUTIONS. And the distribution of the vector $\boldsymbol{A}_n = (A_1(n), A_2(n), \ldots)$ is given by Donnelly and Tavaré [9] as

$$\mathbb{P}[K_n = k, A_i(n) = n_i, 1, 2, \ldots, k] = \frac{\theta^k (n-1)!}{\theta_{(n)} n_k (n_k + n_{k-1}) \cdots (n_k + n_{k-1} + \cdots + n_2)}. \quad (5)$$

Thus the conditional distribution of $\boldsymbol{C}_n$, given $K_n = k$, is

$$\mathbb{P}[\boldsymbol{C}_n = \boldsymbol{a}_n \mid K_n = k] = \frac{n!}{S(n,k) \prod_j j^{a_j} a_j!}, \quad (6)$$

and the number of singleton species has the distribution

$$\mathbb{P}[C_1(n) = a] = \frac{\theta^a}{a!} \sum_{j=0}^{n-a} (-1)^j \frac{\theta^j}{j!} \frac{(n+1-a-j)_{(a+j)}}{(n+\theta-a-j)_{(a+j)}}. \quad (7)$$

The joint factorial moments* of $\boldsymbol{C}_n$, of arbitrary order, are

$$\mathbb{E} \prod_{j=1}^{n} C_j(n)_{[r_j]} = \frac{n!}{m!} \frac{\theta_{(m)}}{\theta_{(n)}} \prod_{j=1}^{n} \left(\frac{\theta}{j}\right)^{r_j}$$

when $m = n - \sum j r_j \geq 0$, and 0 when $m < 0$ (Watterson [33]); here we define $x_{[r]} = x(x-1)\cdots(x-r+1)$ for $r = 0, 1, 2, \ldots$. Also, the mean and variance of K_n are

$$\mathbb{E}(K_n) = \sum_{i=0}^{n-1} \frac{\theta}{\theta + i},$$

$$\mathrm{Var}(K_n) = \sum_{i=1}^{n-1} \frac{\theta i}{(\theta + i)^2}.$$

These are the core consequences of the probability model assumed for the ESF.

PROPERTIES AND CHARACTERIZATIONS

The ESF is a member of the exponential family* of distributions. The complete sufficient statistic for θ is K_n, and its maximum likelihood estimator* $\hat{\theta}$ is given implicitly from (4) as the solution of the equation $\sum_{i=0}^{n-1} \hat{\theta}/(\hat{\theta} + i) = K_n$. The only functions of θ admitting unbiased* estimation are linear combinations of expressions of the form $[(i+\theta)(j+\theta)\cdots(m+\theta)]^{-1}$, where $i, j, \ldots, m$ are integers with $1 \leq i < j < \cdots < m \leq n-1$.

Let μ_n denote the distribution of the partition vector $\boldsymbol{C}_n$ when sampling from the species model in (1). We say the sample has the *species deletion property* if, when an animal is taken at random from the sample and it is observed that in all there are r animals of this species in the sample, then the partition distribution of the remaining $n - r$ animals is μ_{n-r}. Kingman [22, 23] shows that the species deletion property holds for the ESF [i.e., when μ_n is given by (3)].

Next we consider the properties of (3) and (5) for two consecutive sample sizes, n and $n+1$. We denote the history of the sample of size n by $\mathcal{H}_n = (\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots, \boldsymbol{A}_n)$, where $\boldsymbol{A}_i$ is the vector describing the species composition of the first i animals. We ask: Given $\mathcal{H}_n$, what is the conditional probability that the next animal will be of a new species? This probability is found from (3) as

$$\mathbb{P}[(n+1)\text{th animal of a new species} \mid \mathcal{H}_n] = \frac{\theta}{n+\theta}. \quad (8)$$

If a given species has been observed m times ($m > 0$) in the sample of n, the conditional probability that the $(n+1)$th animal will be of this species is

$$\mathbb{P}[(n+1)\text{th animal is of a particular species seen } m \text{ times} \mid \mathcal{H}_n] = \frac{m}{n+\theta}. \quad (9)$$

The probabilities (8) and (9) may be used to generate the process $\boldsymbol{A}_n$, $n = 1, 2, \ldots$, by a sequential urn scheme, starting from $\boldsymbol{A}_1 = (1)$. This scheme is a special case of an urn model* considered, in a nonspecies context, by Blackwell and MacQueen [6], who used it to

discuss properties of sampling from a Dirichlet process*. Hoppe [18] was the first to exploit this urn model in the context of statistical genetics* to obtain properties of the distribution (3).

The species deletion property and the law-of-succession (*see* LAPLACE'S LAW OF SUCCESSION) formulas (8) and (9) may be used to characterize the ESF in the context of sampling from the model (1):

1. (Kingman [22, 23].) If the species deletion property holds, then the vector $\boldsymbol{C}_n$ has distribution μ_n given by the ESF.
2. The Law of Succession. Suppose that the sample history $\mathcal{H}_n$ is given. If the conditional probability that the next animal is of a new species depends only on n, then this probability must be of the form $\theta/(\theta + n)$ for some nonnegative constant θ [8]. If, further, the conditional probability that this animal is of a specific species seen m times in the sample depends only on m (the sufficientness principle of Johnson [19]), then the species partition probability is given by the ESF [35].

RELATIONS WITH OTHER DISTRIBUTIONS

The ESF can be derived by a conditioning argument from the logarithmic series distribution* [33] as well as from the Poisson*. For the latter, suppose that $Z_1, Z_2 \ldots$ are independent Poisson random variables with $\mathbb{E}[Z_j] = \theta/j$. Then

$$(C_1, \ldots, C_n) =_d \left(Z_1, \ldots, Z_n \;\middle|\; \sum_{j=1}^{n} jZ_j = n \right),$$

where $=_d$ denotes equality in distribution.

Another representation, called the *Feller coupling*, is useful for deriving asymptotic results for the ESF [4]. Let ξ_i, $i \geq 1$, be independent Bernoulli random variables satisfying $\mathbb{P}[\xi_i = 1] = \theta/(\theta + i - 1)$, and let $C_j(n)$ be the number of spacings* of length j between the 1's in the sequence $\xi_1\xi_2 \ldots \xi_n 1$. Then the distribution of the vector $\boldsymbol{C}_n$ is the ESF. Further, if Z_j is the number of spacings of length j in the infinite sequence $\xi_1\xi_2 \ldots$, then the Z_j are independent Poisson random variables with mean $\mathbb{E}[Z_j] = \theta/j$.

The distribution of the vector $\boldsymbol{P} = (P_1, P_2, \ldots)$ determined by (2) is known as the *GEM distribution*. It is named after McCloskey [25] and Engen [10], who introduced it in the context of ecology, and Griffiths [15], who first noted its importance to genetics. The GEM distribution is a residual allocation model [27], that is, a model of the form (2) where $W_1, W_2, \ldots$ are independent. It is the only such model with identically distributed residual fractions for which the size-biased permutation has the same distribution as $\boldsymbol{P}$ (for a definition of a size-biased permutation, and a proof of this assertion, see refs. [25] and [10]).

The decreasing order statistics* $(P_{(1)}, P_{(2)}, \ldots)$ of $\boldsymbol{P}$ have the Poisson–Dirichlet distribution with parameter θ [20]. The GEM is the size-biased permutation of the Poisson–Dirichlet [27, 24].

The ESF is a particular case of the *Pitman sampling formula* [28, 30], which gives the probability of a species partition $\boldsymbol{C}_n = \boldsymbol{a}_n$ of n animals as

$$\mathbb{P}[C_n = a_n, K_n = k] = \frac{n!\,(\theta + \alpha)(\theta + 2\alpha)\cdots[\theta + (k-1)\alpha]}{(\theta + 1)_{(n-1)}} \times \prod_{j=1}^{n} \left(\frac{(1-\alpha)_{(j-1)}}{j!} \right)^{a_j} \frac{1}{a_j!}.$$

Since we are considering only the infinitely-many-species case, we employ the restrictions $0 \leq \alpha < 1$, $\theta > -\alpha$. The ESF is then the particular case of the Pitman sampling formula when $\alpha = 0$. The Pitman distribution has several important properties, of which we note here one. Suppose in the residual allocation model (2) we no longer assume that $W_1, W_2, \ldots$ are identically distributed. Then the most general distribution of W_i for which the distribution of $(P_1, P_2, P_3, \ldots)$ is invariant under size-biased sampling is that for which W_i has probability density proportional to $w^{-\alpha}(1-w)^{\theta+i\alpha-1}$ (cf. ref. [29]). This model for (2) yields the Pitman sampling formula.

It follows from (7) and the method of moments* that random variables $\boldsymbol{C}_n$ with the ESF (3) satisfy, for each fixed b,

$$(C_1(n), \ldots, C_b(n)) \Rightarrow (Z_1, \ldots, Z_b) \qquad (10)$$

as $n \to \infty$, with $\Rightarrow$ denoting convergence in distribution. Rates of convergence in the total-variation* metric are given by Arratia et al. [4].

The approximation in (10) covers the case of species represented a small number of times. A functional central limit theorem* is available for the number of species represented at most n^t times, for $0 < t \leq 1$ [16]. In particular, the number K_n of species in the sample has asymptotically a normal distribution with mean and variance $\theta \log n$.

It follows directly from the strong law of large numbers* that the proportions A_n/n converge almost surely as $n \to \infty$, and the limit has the GEM distribution with parameter θ. The decreasing-order statistics of A_n/n converge almost surely to the Poisson–Dirichlet distribution with parameter θ [20].

APPLICATIONS

We summarize applications of the ESF in various different areas. A more extensive review of these topics may be found in Tavaré and Ewens [32].

The ESF provides the null-hypothesis distribution of allele frequencies for the "non-Darwinian" theory of evolution in population genetics. The parameter θ in (3) depends on the population size, the mutation rate, and details of the evolutionary model, all usually unknown. However, the conditional distribution (6), being independent of θ, can be used for testing this theory even when these quantities are unknown. For details see Watterson [34].

In the context of Bayesian statistics, Antoniak [2] showed that the ESF gives the distribution of the partition of a sample from a Dirichlet process prior. Ferguson et al. [13] and Sethuraman [31] give recent developments.

Equation (3) arises in a number of combinatorial contexts. First, as noted earlier, the case $\theta = 1$ gives the distribution of the number and lengths of the cycles in a uniform random permutation of n objects. If the permutation is chosen with probability proportional to θ^t, where t equals the number of cycles, then this distribution is given by (3). Second, suppose a random mapping of $(1, 2, \ldots, N)$ to $(1, 2, \ldots, N)$ is made, each mapping having probability N^{-N}. Then in the limit as $N \to \infty$, the images of the components of this mapping in the set $(1, 2, \ldots, n)$ have the distribution (3) with $\theta = \frac{1}{2}$ [21, 1]. Further combinatorial structures where (3) arises, including factorizations of polynomials over a finite field, are described by Arratia and Tavaré [3].

In ecology, the species number and size allocation distribution where species do not interact is of some interest. This corresponds to the independent sampling property of the ESF; its characterization shows that this distribution provides the required partition. The same property is also used in physics [26, 17].

Bartholomew [5] describes a simple model of the spread of news throughout a population, where each individual hears the news first either from a source (e.g., a radio station) or from someone who already knows the news. Individuals can be grouped into components, each component consisting of one individual who first heard the news from the source together with all those individuals who first heard it through some chain deriving from this person. The distribution of component number and sizes is given by (3), where θ is the ratio of the rates at which individuals in the population hear the news from the source and from other individuals.

References

[1] Aldous, D. J. (1985). Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII—1983*, Lecture Notes in Mathematics 1117. P. L. Hennequin, ed., Springer-Verlag, Berlin, pp. 2–198.

[2] Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.*, **2**, 1152–1174.

[3] Arratia, R. A. and Tavaré, S. (1994). Independent process approximations for random combinatorial structures. *Adv. Math.*, **104**, 90–154.

[4] Arratia, R. A., Barbour, A. D., and Tavaré, S. (1992). Poisson process approximations for the Ewens sampling formula. *Ann. Appl. Probab.*, **2**, 519–535.

[5] Bartholomew, D. J. (1973). *Stochastic Models for Social Processes*, 2nd ed. Wiley, London.

[6] Blackwell, D. and MacQueen, J. B. (1973). Ferguson distributions via Pólya urn schemes. *Ann. Statist.*, **1**, 353–355.

[7] Cauchy, A. (1905). *Oeuvres Complètes*, II Ser., vol. 1. Gauthier-Villars, Paris.

[8] Donnelly, P. (1986). Partition structures, Polya urns, the Ewens sampling formula, and the ages of alleles. *Theor. Population Biol.*, **30**, 271–288.

[9] Donnelly, P. and Tavaré, S. (1986). The ages of alleles and a coalescent. *Adv. Appl. Probab.*, **18**, 1–19.

[10] Engen, S. (1975). A note on the geometric series as a species frequency model. *Biometrika*, **62**, 697–699.

[11] Engen, S. (1978). *Stochastic Abundance Models with Emphasis on Biological Communities and Species Diversity*. Chapman and Hall, London.

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

[13] Ferguson, T. S., Phadia, E. G., and Tiwari, R. C. (1992). Bayesian nonparametric inference. In *Current Issues in Statistical Inference: Essays in Honor of D. Basu*, M. Ghosh and P. K. Patnak, eds., IMS Lecture Notes–Monograph Ser. 17. Institute of Mathematical Statistics, Hayward, Calif., pp. 127–150.

[14] Fisher, R. A., Corbet, A. S., and Williams, C. B. (1943). The relation between the number of species and the number of individuals in a random sample from an animal population. *J. Animal Ecol.*, **12**, 42–58.

[15] Griffiths, R. C. (1980). Unpublished notes. Monash Univ., Melbourne, Australia.

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

[17] Higgs, P. G. (1995). Frequency distributions in population genetics parallel those in statistical physics. *Phys. Rev. E*, **51**, 95–101.

[18] Hoppe, F. M. (1987). The sampling theory of neutral alleles and an urn model in population genetics. *J. Math. Biol.*, **25**, 123–159.

[19] Johnson, W. E. (1932). *Logic, Part III: The Logical Foundations of Science*. Cambridge University Press.

[20] Kingman, J. F. C. (1975). Random discrete distributions. *J. R. Statist. Soc. B*, **37**, 1–22.

[21] Kingman, J. F. C. (1977). The population structure associated with the Ewens sampling formula. *Theor. Population Biol.*, **11**, 274–283.

[22] Kingman, J. F. C. (1978). Random partitions in population genetics. *Proc. R. Soc. London A*, **361**, 1–20.

[23] Kingman, J. F. C. (1978). The representation of partition structures. *J. London Math. Soc.*, **18**, 374–380.

[24] Kingman, J. F. C. (1993). *Poisson Processes*. Clarendon Press, Oxford.

[25] McCloskey, J. W. (1965). A model for the distribution of individuals by species in an environment. Ph.D. thesis, Michigan State University.

[26] Mekjian, A. Z. (1991). Cluster distributions in physics and genetic diversity. *Phys. Rev. A*, **44**, 8361–8374.

[27] Patil, G. P. and Taillie, C. (1977). Diversity as a concept and its implications for random communities. *Bull. Int. Statist. Inst.*, **XLVII**, 497–515.

[28] Pitman, J. (1992). The two-parameter generalization of Ewens' random partition structure. *Tech. Rep. 345*, Department of Statistics, University of California, Berkeley.

[29] Pitman, J. (1996). Random discrete distributions invariant under size-biased permutation. *Ann. Appl. Probab.*, **28**, 525–539.

[30] Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probab. Theory and Related Fields*, **102**, 145–158.

[31] Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Stat. Sinica*, **4**, 639–650.

[32] Tavaré, S. and Ewens, W. J. (1997). The Ewens multivariate distribution. In *Discrete Multivariate Distributions*, N. L. Johnson, S. Kotz, and N. Balakrishnan, authors. Wiley, New York, Chapter 41.

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

[34] Watterson, G. A. (1978). The homozygosity test of neutrality. *Genetics*, **88**, 405–417.

[35] Zabell, S. L. (1998). The continuum of inductive methods revisited. Pittsburgh–Konstanz Series in the History and Philosophy of Science. To appear.

(CLASSICAL DISCRETE DISTRIBUTIONS, GENERALIZATIONS OF
COMBINATORICS
FACTORIAL SERIES DISTRIBUTIONS
GENETICS, STATISTICS IN
LAPLACE'S LAW OF SUCCESSION
SPACINGS
STIRLING DISTRIBUTIONS)

W. J. EWENS
S. TAVARÉ

EXPERT SYSTEMS, PROBABILISTIC

Expert systems are computer programs meant to assist or replace humans performing compli-