Statistical Science
2016, Vol. 31, No. 1, 23–26
DOI: 10.1214/15-STS536
Main article DOI: 10.1214/15-STS529

Two Early Contributions to the Ewens Saga

Peter McCullagh

Abstract. The mixture model devised by Fisher, Corbet and Williams [*Journal of Animal Ecology* **12** (1943) 42–58] for species sampling and the sequential prediction approach pioneered by Good [*Biometrika* **40** (1953) 237–264] and Good and Toulmin [*Biometrika* **43** (1956) 45–63] are both closely related to the Ewens sampling formula. Fisher's two-parameter joint distribution for the species counts includes the Ewens distribution as the conditional distribution given the sample size. The log-series model, as it is known in the ecological literature, is closely related to a Poisson process model devised by Arratia, Barbour and Tavaré [*Ann. Appl. Probab.* **2** (1992) 519–535]. Oddly, despite its advantages for statistical inference, Fisher does not mention the conditional distribution. Likewise, athough Good (1953) pioneered the sequential prediction approach, neither he nor Toulmin discovered the Ewens process in a form equivalent to the modern-day Chinese restaurant process.

Key words and phrases: Chinese restaurant process, Poisson process, species richness, species sampling.

Crane is to be commended for his survey of the diverse areas of scientific work in which the Ewens sampling formula has arisen. It is an impressive list stretching from literary studies to population genetics and probabilistic number theory. The fundamental mathematical object in all of this work is a partition—originally an integer partition but preferably a set partition—which splits the population units into disjoint subsets called blocks or clusters and does the same thing to the sample units. Crane makes a strong case that the Ewens process is to random partitions or clusterings as the Gaussian process is to a random sequence of real numbers, or the Poisson process is to a random series of events in time or space. I agree. It is one of a small number of processes that deserves to be a central part of the statistical curriculum.

Fisher, Corbet and Williams (1943) appears to be one of the first studies of the statistical relation between the number of specimens and the number of species in typical ecological samples. In this setting, the multiplicity vector m with components m_r records the number of species for which exactly r specimens occur in the sample, so $m_. = \sum_{r>0} m_r$ is the total species count, and $N = \sum_r r m_r$ is the specimen count.

Although Fisher's paper is a citation classic in the ecological literature, and the approach appears to be simple and well understood, some parts of his argument are not straightforward and other parts are not correct. Fisher begins with the assumption that the specimen counts for a single species are distributed according to the Poisson distribution with parameter λ, and the species-specific λ-values are distributed according to a Poisson process with mean measure G on $(0, \infty)$. It follows that the joint intensity-count distribution is that of a Poisson process $Z \sim \mathrm{PP}(\mu)$ with mean measure

$$\mu(d\lambda, r) = \frac{e^{-\lambda}\lambda^r}{r!} G(d\lambda)$$

at (λ, r) in the product space. The projected marginal process on counts, the frequency of frequencies, $m_r = \#\{\lambda : (\lambda, r) \in Z\}$ is also Poisson, and if G is proportional to the gamma distribution, the marginal measure at $r \geq 0$ is

$$\mu_r = E(m_r) = \theta \frac{\Gamma(r + \nu)\eta^r}{r!}, \tag{1}$$

Peter McCullagh is Professor, Department of Statistics, University of Chicago, Chicago, Illinois 60637, USA (e-mail: pmcc@galton.uchicago.edu).

proportional to the negative binomial series for some $0 < \eta < 1$. In ecological work, the observation is not the marginal process, but its restriction to $r > 0$, excluding m_0.

To understand better the interpretation of these parameters, it is helpful to consider the effect of increasing the sampling effort by the factor $t > 0$, for example, by increasing the number of traps or the observation time. All things being equal, the effect on each species intensity is $\lambda \mapsto t\lambda$, so the transformed measure is $G_t(A) = G(t^{-1}A)$ for subsets $A \subset \mathcal{R}^+$. The total species mass is unaltered. The effect on the marginal mean measure (1) is to transform η to η_t multiplicatively on the odds scale:

$$\frac{\eta_t}{1-\eta_t} = t\frac{\eta}{1-\eta}, \tag{2}$$

leaving θ fixed. My guess is that Fisher was aware of the implication, $\eta_t = t/(t+\gamma)$ for some constant $\gamma > 0$ independent of t, but this equation does not appear in his paper.

On the basis of empirical evidence derived from Corbet's series on Malayan butterflies, Fisher concluded that ν must be close to zero; the limit value $\nu = 0$ implies that $\mu_0 = \infty$ and $\mu_r = \theta\eta^r/r$ proportional to the coefficients in the expansion of $-\log(1-\eta)$. Conveniently enough, Fisher's limiting log-series model, $m_r \sim \mathrm{Po}(\mu_r)$ with infinitely many independent components, is log-linear

$$\log\mu_r = \log\theta + r\log\eta - \log r.$$

It is a generalized linear model with canonical parameter $(\log\theta, \log\eta)$, offset $-\log r$, and minimal sufficient statistic $(m_., N)$ with expected value

$$\begin{aligned} &E(N) = \theta\eta/(1-\eta), \\ &E(m_.) = -\theta\log(1-\eta) = \theta\log\bigl(1 + E(N)/\theta\bigr). \end{aligned} \tag{3}$$

Fisher computed the maximum-likelihood estimate by solving the simultaneous nonlinear equation

$$N = \hat\theta\hat\eta/(1-\hat\eta), \quad m_. = -\hat\theta\log(1-\hat\eta).$$

In the pre-computer era, he also provided tables to assist in its solution.

For each fixed θ, the statistic N is complete and sufficient for η, so Fisher's two-parameter model has a Neyman structure (Lehmann, 1986, Section 4.3). Given $N = n$, the multiplicity vector m is a random partition of the integer n. By sufficiency, the conditional distribution depends on θ alone; it is the Ewens sampling formula with parameter θ. Given his earlier writings on 2 × 2 tables in *Statistical Methods for Research Workers* (1935, Section 4) and his 1934 paper on location-scale models, it is strange that Fisher does not mention the conditional distribution. Presumably it did not occur to him then or subsequently. But one Fisherian passage is worth quoting: *The quantity θ is independent of the size of sample and is proportional to the number of species of the group considered, at any chosen level of abundance, relative to the means of capture employed. Values of θ from different samples [...] may be compared as a measure of richness in species.* Fisher's expressions (3) for the moments do not justify the first part of his statement, so presumably what he had in mind was the argument leading to (2), undoubtedly obvious to Fisher if not to most readers, that increased sampling effort leaves θ fixed but increases η in an inverse-linear manner.

Fisher's paper concludes with a discussion of maximum-likelihood estimation and the computation of standard errors, both of which would have been simpler using the conditional likelihood. For Williams's Macrolepidoptera series at Harpenden, the counts $N = 15{,}609, m_. = 240$ yield $\hat\theta = 40.248$ and $\hat\eta = 0.9974281$; the conditional mle of θ based on the Ewens model is 40.146. The unconditional and conditional standard errors for $\hat\theta$ are 2.85 and 2.84 respectively. The observed frequencies are in remarkably good agreement with the fitted series, and the fit is not improved by taking $\nu \neq 0$. The proximity of $\hat\eta$ to the upper boundary is consistent with the asymptotic behavior of the Ewens process (Arratia, Barbour and Tavaré, 1992), so $\eta = 1$ is the only correct value in the limit.

For Fisher's two-parameter model, the asymptotic variance of $\hat\theta$, as given by the eponymous inverse information matrix, is

$$\frac{\hat\theta^2}{m_. - N(1-\hat\eta)} = \frac{\hat\theta^2(N+\hat\theta)}{m_.(N+\hat\theta) - N\hat\theta};$$

the numerical value is 8.105. Using a line of argument that is flawed in places, Fisher deduced incorrectly that $\mathrm{var}(m_.) \simeq \theta\log 2$ instead of $\theta\log(1 + N/\theta)$. As a result, the formula given for the variance of $\hat\theta$ is incorrect, and the reported value (1.1251) is too small by the approximate factor $\log(1+N/\theta)/\log(2) = 8.60$.

Ten years later, Good looked at the problem of estimating the population relative frequency q_r of a species having r representatives in the sample, whose size n is fixed by design. The two authors have very different styles. Fisher's four-page contribution is terse to the point of obscurity; Good's 25-page tracts are discursive to the point of distraction. Fisher embraces parametric assumptions; Good recognizes the need for smoothing, but he avoids parametric assumptions—even when they might be helpful.

Avoiding all assumptions about the behavior of the expected multiplicities, $\mu_{n,r} = E(m_r | N = n)$, Good (1953) concluded that the posterior expected frequency is

$$E(q_r|\text{data}) = \frac{r+1}{n+1}\frac{\mu_{r+1,n+1}}{\mu_{r,n}}. \tag{4}$$

Subsequently, Good and Toulmin (1956) looked at the problem of sample extension, in an effort to determine the conditional distribution of the number of new species that occur in an extended sample. Their approach puts the emphasis on prediction, where it properly belongs, rather than on distribution fitting and parameter estimation. It may be viewed as the first attempt to express an infinitely exchangeable partition process dynamically using the conditional distribution given the current configuration. If they had adapted Fisher's two-parameter model, they might have succeeded in developing a Chinese-restaurant description. As it stands, their analysis is unavoidably complicated because of the deliberate avoidance of parametric assumptions.

For a point-process model in which n is not fixed, Good's predictive ratio is replaced by the Papangelou conditional intensity ratio with $\mu_r \propto \eta^r/r$, which yields

$$E(q_r|\text{data}) \propto \begin{cases} r\eta, & r \geq 1; \\ \theta\eta, & r = 0, \end{cases}$$

for Fisher's model. The value for $r = 0$ is the combined intensity for all unrepresented species. Had Good or Good and Toulmin taken a more positive view of parametric models, they could easily have arrived at the simpler expression

$$E(q_r|\text{data}) = \frac{r}{n+\theta},$$

leading to $q_0 = \theta/(n+\theta)$, which is exact for Fisher's model. But their determination to develop the story without the benefit of parametric smoothing led them elsewhere. Recognizing that the posterior expected value of q_r is the same as the conditional probability that the next specimen belongs to that block, we can speculate on how the subject might have developed differently.

It is fair to say that Fisher almost discovered the Ewens sampling formula. He had it in his grasp. After all, he had only to compute the conditional distribution given the sample size, a task that was both statistically natural and, for him, mathematically trivial. But he did not. And even had he done so, the conditional distribution would not necessarily have led quickly to the Ewens *process* as we know it today. It is also fair to say that Good *should* have discovered the process. He also had it in his grasp, in the sense that he was aware of Fisher's work, he was asking the right questions, and his analysis was correct. But his refusal to consider Fisherian parametric smoothing was a critical blinker. In the end, Fisher did not discover the sampling formula, and Good discovered neither the formula nor the process.

In closing, it is of some interest to examine Williams's Macrolepidoptera data for the years 1933–36 from the perspective of the Ewens process. For the year 1933, 178 species were observed in a sample of 3540 specimens, yielding $\hat{\theta}_1 = 39.355$ with standard error 3.34. The data given by Williams in Table 4 is not sufficient to determine the additional species numbers for each of the following three years, but, for the three years combined, a further set of 62 species was recorded among 12,069 specimens. In the Ewens process, this additional species count is a sum of independent Bernoulli variables with mean 58.06 and variance 57.73, so the observed value is only 0.52 standard deviations above what is predicted under the model of temporal homogeneity using the fitted 1933 value $\hat{\theta}_1$.

Given the observed data for the year 1933, the Ewens conditional likelihood yields an estimated richness parameter $\hat{\theta}_2 = 42.042$ with standard error 5.36 for the period 1934–36. Since these estimators are statistically independent, at least asymptotically, we may compute a standardized difference in the usual way, giving the value $T = 2.687/6.314 = 0.425$ as a test for temporal homogeneity. The overall maximum likelihood estimate, assuming temporal homogeneity for the combined sample, is $\hat{\theta} = 40.146$, and the likelihood ratio statistic is 0.184, in good agreement with the Wald statistic $T^2 = 0.181$. Although the specimen count for 1935 was more than twice the count in any other year, there is no evidence for a change in the relative composition of the Macrolepidoptera population at Harpenden over these four years.

REFERENCES

ARRATIA, R., BARBOUR, A. D. and TAVARÉ, S. (1992). Poisson process approximations for the Ewens sampling formula. *Ann. Appl. Probab.* **2** 519–535. MR1177897

FISHER, R. A. (1934). Two new properties of maximum likelihood. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **144** 285–307.

FISHER, R. A. (1935). *Statistical Methods for Research Workers*. Oliver & Boyd, Edinburgh.

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 of an animal population. *Journal of Animal Ecology* **12** 42–58.

GOOD, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika* **40** 237–264. MR0061330

GOOD, I. J. and TOULMIN, G. H. (1956). The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika* **43** 45–63. MR0077039

LEHMANN, E. L. (1986). *Testing Statistical Hypotheses*, 2nd ed. Wiley, New York. MR0852406