Bernoulli **29**(2), 2023, 1166–1194
https://doi.org/10.3150/22-BEJ1494

Fisher's measure of variability in repeated samples

POLY H. DA SILVA[1,a], ARASH JAMSHIDPEY[2,b], PETER MCCULLAGH[3,c] and SIMON TAVARÉ[1,4,d]

[1]*Department of Statistics, Columbia University, 1255 Amsterdam Ave, New York, NY 10027, USA,* [a]*phd2120@columbia.edu*
[2]*Department of Mathematics, Columbia University, 2990 Broadway, New York, NY 10027, USA,* [b]*aj2963@columbia.edu*
[3]*Department of Statistics, University of Chicago, 5747 S Ellis Ave, Chicago, IL 60637, USA,* [c]*pmcc@galton.uchicago.edu*
[4]*Irving Institute for Cancer Dynamics, Columbia University, Schermerhorn Hall, Suite 601, 1190 Amsterdam Avenue, New York, NY 10027, USA,* [d]*st3193@columbia.edu*

Fisher (1943) claimed that the expected value of the sample variance of the number of species found in large samples, each of n specimens taken from the same population, is asymptotically $\theta \log 2$. This is at odds with the value $\theta \log n$ obtained directly from the Ewens Sampling Formula (ESF), where θ specifies the rate at which new species are found. To resolve this apparent contradiction, we assume the species frequency spectrum in the population is determined by the ESF and that the samples are disjoint subsets drawn sequentially from this single population. We find an explicit formula for the required expected value for p samples of arbitrary size; in the limit of large equally-sized samples, it indeed has the value $\theta \log 2$. We obtain limit theorems for the sample variance of p samples of size n under various limiting regimes as p, n or both tend to ∞. We discuss further the behavior of the number of species present in all samples, and revisit Fisher's log-series distribution as the limiting distribution of the number of specimens observed in typical species in a future, large sample.

Keywords: Ewens Sampling Formula; log-series model; sequential sampling; exchangeability; Chinese Restaurant Process; Poisson approximation

1. Introduction

Fisher, Corbet and Williams [14] studied the relationship between the number of species and the number of specimens found in typical ecological samples, illustrating their analysis with data from a Microlepidoptera sample from England, and another from Malayan butterflies. Despite its popularity in the ecological literature, there remain some statistical aspects of their modeling that might benefit from further investigation, and we provide one view here.

It is convenient to use the notation $C = (C_1, C_2, \ldots)$ to denote species counts, C_i denoting the number of species observed i times in the sample; the total number of species observed is

$$S = C_1 + C_2 + \cdots$$

and the number of specimens is

$$N = C_1 + 2C_2 + 3C_3 + \cdots .$$

The literature now describes many models in which either, or both, of S and N are viewed as random. For example, [29] discusses five versions of log-series models, and describes the statistical issues involved in estimating parameters in the models for S and N. See also [19,24] and [10, Chapter 3]. We

stay with Fisher's [13] original limiting log-series model in which the counts $C_i, i \geq 1$ are independent Poisson random variables with means given by

$$m_i := \mathbb{E}C_i = \theta \frac{\eta^i}{i} \tag{1}$$

for parameters $\eta \in (0,1)$ and $\theta > 0$. N then has a negative binomial distribution with

$$\mathbb{P}(N = n) = \binom{\theta + n - 1}{n} (1-\eta)^\theta \eta^n, n = 0, 1, \ldots \tag{2}$$

and mean

$$\mathbb{E}N = \frac{\theta\eta}{1-\eta},$$

while S has a Poisson distribution with mean

$$\mathbb{E}S = -\theta \log(1-\eta).$$

Fisher was concerned with estimation of the parameters θ and η based on the observed values of N and S. He showed, inter alia, that the maximum likelihood (and moment) estimates solved the simultaneous equations

$$N = \frac{\hat{\theta}\hat{\eta}}{1-\hat{\eta}}, \quad S = -\hat{\theta}\log(1-\hat{\eta}),$$

and concluded with a discussion of standard errors of the estimates when the observed number of specimens, N, is large.

Fisher was particularly concerned with the variance expected in estimated values of θ in samples taken from the same population of species. To this end, let $S_1, \ldots, S_p$ denote the values of S observed in p disjoint samples, and define the sample variance by

$$V_p = \frac{1}{p(p-1)} \sum_{i<j} (S_i - S_j)^2 \tag{3}$$

If, in fact, the samples were mutually independent and approximately identically distributed with the same, large value of N, then $\mathbb{E}V_p = \operatorname{var} S = \theta \log(1 + \mathbb{E}(N)/\theta) \approx \theta \log N$. Fisher, however, claimed that $\mathbb{E}V_p \approx \theta \log 2$.

The difference between these formulae has implications for estimates of θ: Fisher's result implies that the expected value of the sample variance of the values of $\hat{\theta}^{(i)} \approx S_i/\log N$ is of order $\theta \log 2/(\log N)^2$, whereas the other calculation gives an estimate of order $\theta/\log N$.

The interpretation of one form of variance over the other is the key to understanding the goal of the calculation by Fisher [13] of the 'variation of S and N in parallel samples'. The discrepancy between the two variance formulae was noted by F.J. Anscombe, who set out to clarify the matter at a meeting with Fisher in April 1947, and in correspondence thereafter. Fisher's derivation is both brief and opaque, and his conclusion as stated is not correct, or is at least open to mis-interpretation. However, a passage taken from [2, Section 6] suggests that what Fisher had in mind was a different notion of variability. According to Anscombe, Fisher's variance formula *is appropriate to a special type of comparison, namely, between estimates of* α [our θ] *... derived from similar nearby traps, where it may be assumed that the individual species have exactly the same relative abundances, and the difference between the catches at any two traps arises solely from Poisson variation in the numbers caught of each species.*

While he appeared to accept Fisher's claim at face value, Anscombe offered no proof. The phrasing of his explanation, which hews closely to passages from his correspondence with Fisher, offers no insight into the relevant paragraph of Fisher's paper.

Watterson [29, pp. 218–219] also recognizes that Fisher's moment calculations indicate that *he had an entirely different version of the logarithmic distribution in mind when he came to consider what would actually appear in a sample trapping*. Version 5 of the log series model in Watterson (1974) offers a lengthy and somewhat complicated interpretation of what Fisher seems to have had in mind. This involves a double limit in which a series is approximated by an integral, which leads to a pseudo-generating function in which probabilities are not necessarily non-negative. From this, Watterson suggests a way to make the argument more rigorous, leading eventually to Fisher's variance formulae in equation 2.43. The argument does not offer a confirmation of Anscombe's interpretation, nor does it shed light on which version is more appropriate for what purpose.

This paper offers a simple sequential rationalization of Fisher's variance formula, which is in line with the quote from Anscombe. More importantly, it does so by exploiting a familiar framework that avoids the technical complications of Watterson's derivation (version 5). However, our derivation sheds absolutely no light on what Fisher had in mind or how he arrived at his formula.

2. A sequential sampling approach

As noted in [24], Fisher did not describe the conditional distributions of C or S given $N = n$, which arguably would have made his analysis more transparent. The conditional distribution of C given $N = n$ is readily found from (1) and (2). For non-negative integers $c_1, c_2, \ldots, c_n$ satisfying $c_1 + 2c_2 + \cdots + nc_n = n$, one obtains

$$\mathbb{P}(C_1 = c_1, \ldots, C_n = c_n | N = n) = \frac{n!}{\theta_{(n)}} \prod_{j=1}^{n} \left(\frac{\theta}{j}\right)^{c_j} \frac{1}{c_j!}, \tag{4}$$

where

$$\theta_{(n)} = \theta(\theta + 1) \cdots (\theta + n - 1), \quad \theta_{(0)} = 1.$$

The distribution in (4) is known as the Ewens Sampling Formula [11] with parameter θ, denoted $\mathrm{ESF}_n(\theta)$ in what follows when the sample size n needs emphasis. It arose originally in population genetics, but is now found in many different settings. For reviews, see [17, Chapter 41], [8], and for basic theory and applications in combinatorics, see [3,28].

The Ewens model provides a natural setting for Fisher's species sampling problem; n, the number of specimens sampled, is deterministic, and S, the number of species observed, is random. Increasing the sample size corresponds to sampling more of the population. We take a sequential view, in which $S_i := S_i(n_i)$ denotes the number of species found in a sample of size n_i forming the ith of p disjoint samples, and ask what can be said about the sample variance of $S_1, \ldots, S_p$. The species counts are evidently not independent, as later samples may well have contained species that have already been found. It is understanding this dependence that leads, among other things, to a reinterpretation of Fisher's result.

The sequential sampling may be thought of in different ways. For example, [12] considers the result of two investigators taking distinct samples of equal size from the same population at the same time, and [5], in a study of genetic variation in tumours, generalizes this scheme to an arbitrary number of samples of any size.

To set the scene, we note that Ewens [11] gave the distribution of $S = S(n)$, the number of species observed in a sample of size n, as

$$\mathbb{P}(S(n) = k) = \frac{\theta^k \left[{n \atop k}\right]}{\theta_{(n)}}, k = 1, \ldots, n, \tag{5}$$

where $\left[{n \atop k}\right]$ denotes the unsigned Stirling number of the first kind. He showed that the maximum likelihood estimator $\hat{\theta}$ of θ is asymptotically given by $\hat{\theta} = S_n/\log n$, and that the variance of this estimator is asymptotically $\theta/\log(n)$. Thus the expected value of the between-sample variance of p *independent* θ estimates is, from (3), asymptotically $\theta/\log n$, as given in the introduction.

For the sequential sampling setting in which p samples of equal size n are taken, the counts S_i are exchangeable, and so the expected value of the sample variance is

$$\mathbb{E}V_{p,n} = \operatorname{var} S_1 - \operatorname{cov}(S_1, S_2).$$

As a consequence of Theorem 1 we show, among other results, that

$$\mathbb{E}V_{p,n} = \sum_{r=n}^{2n-1} \frac{\theta}{\theta + r} - \frac{\theta n!}{\theta_{(2n)}} \sum_{r=1}^{n} \frac{1}{r} \frac{\theta_{(2n-r)}}{(n-r)!} \sum_{i=n-r}^{n-1} \frac{\theta}{\theta + i}. \tag{6}$$

It follows that, as $n \to \infty$,

$$\mathbb{E}V_{p,n} \to \theta \log 2,$$

as shown in [12] for $p = 2$ and [5] for arbitrary p, and confirming Fisher's approximate formula for the expected sample variance of estimators of θ.

The paper is laid out as follows. In Section 3 we derive a formula for the covariance of the species counts from two samples, of size m and n respectively, and describe the asymptotics in the regime in which $n/m \to \beta$ as $n, m \to \infty$. Section 4 sets the scene for the multiple sample case by drawing on notions of exchangeability. This provides the ingredients for studying, in Section 5, the asymptotic distributional behavior of $V_{p,n}$ when p sequential samples of size n are taken, and one or both of p and n tend to infinity in a prescribed way. Section 6 takes a further look at the number of species found in all the samples, and Section 7 comes full circle to describe another appearance on Fisher's log-series distribution.

3. Results for two sequential samples

We begin by reminding the reader about the Chinese Restaurant Process, described here as a model for sampling specimens from different species, a coupling that may be used to generate observations having the distribution of $\mathrm{ESF}_n(\theta)$ for any value of n. To this end let $\xi_1, \xi_2, \ldots$ be a sequence of independent Bernoulli random variables satisfying

$$\mathbb{P}(\xi_i = 1) = \frac{\theta}{\theta + i - 1} = 1 - \mathbb{P}(\xi_i = 0).$$

The parameter $\theta \in (0, \infty)$ represents the rate at which new species are discovered, and $\xi_i = 1$ means that specimen i is from a new species, while $\xi_i = 0$ means that specimen i is of a species that has already been sampled. To complete the description of the dynamics, we use auxiliary randomization: we assume that if an old species is sampled, that species is chosen at random from the existing species

in proportion to the number of that species already observed. The joint distribution of the counts $C_j(n)$ of species represented by j specimens is $\mathrm{ESF}_n(\theta)$, and $S(n) = \sum_{i=1}^n \xi_i$, so that

$$\mathbb{E}S(n) = \sum_{i=1}^{n} \frac{\theta}{\theta + i - 1}, \quad \operatorname{var} S(n) = \sum_{i=1}^{n} \frac{\theta(i-1)}{(\theta + i - 1)^2}. \tag{7}$$

3.1. The expected number of species shared by two samples

We address first a seemingly less general problem, namely the case of two samples, the first composed of $n_1 = m$ specimens, labelled 1,2,...,m, the second composed of $n_2 = n$ specimens, labelled $m+1, m+2, \ldots, m+n$, chosen according to the model described above. The first m specimens produce species counts of $C_1(m), C_2(m), \ldots, C_m(m)$, and a total of $S_1 = S_1(m) := \sum_{j=1}^m \xi_j = C_1(m) + \cdots + C_m(m)$ species; the distribution of S_1 is given by (5) with n there replaced by m. The second sample is collected one specimen at a time, each resulting in either a species that has been found in the first m specimens, or a new species. It is the interaction between the species found in the first m specimens, and the number of those species also found in the second sample that forms the basis of our results. Our approach is in the spirit of [20,21] and [26] but with a different focus.

To analyse this interaction we need some notation. We denote by T_{mn} the number of species found in the second sample that were not found in the first sample, and by K_{mn} the number of species found in the second sample that were also present in the first sample. Clearly,

$$T_{mn} = \sum_{j=m+1}^{m+n} \xi_j,$$

while the number of species $S_2 = S_2(n)$ in the second sample is given by

$$S_2 = K_{mn} + T_{mn}.$$

We will see later that the distribution of S_2 is that of a typical sample of size n, given in (5), so that

$$\mathbb{E}S_2 = \sum_{j=0}^{n-1} \frac{\theta}{\theta + j} = \mathbb{E}K_{mn} + \mathbb{E}T_{mn} = \mathbb{E}K_{mn} + \sum_{j=m}^{m+n-1} \frac{\theta}{\theta + j}.$$

It follows that

$$\mathbb{E}K_{mn} = m\theta \sum_{j=0}^{n-1} \frac{1}{(\theta + j)(\theta + m + j)}. \tag{8}$$

We note that $\mathbb{E}K_{mn} = \mathbb{E}K_{nm}$, as can be verified by interchanging the roles of m and n in (8). An alternative derivation is a consequence of the preliminary results given in Section A.1 to establish Theorem 1; see Lemma 6.

3.2. The covariance of S_1 and S_2

The main result of this section is contained in Theorem 1. Many of the details of the proof are relegated to Section A.1 of the Appendix.

Theorem 1.

$$\operatorname{cov}(S_1(m), S_2(n)) = \operatorname{var} S_1(m) - \sum_{r=n}^{n+m-1} \frac{\theta}{\theta + r} + \frac{\theta m!}{\theta_{(m+n)}} \sum_{r=1}^{m} \frac{1}{r} \frac{\theta_{(m+n-r)}}{(m-r)!} \sum_{i=m-r}^{m-1} \frac{\theta}{\theta + i}. \tag{9}$$

Proof. Notice that $\operatorname{cov}(S_1, S_2) = \operatorname{cov}(S_1, K_{mn} + T_{mn}) = \operatorname{cov}(S_1, K_{mn})$ since S_1 and T_{mn} are independent. From (40),

$$\mathbb{E}S_1 \, \mathbb{E}K_{mn} = (\mathbb{E}S_1)^2 - \mathbb{E}S_1 \, \frac{m!\theta}{\theta_{(m+n)}} \sum_{r=1}^{m} \frac{1}{r} \frac{\theta_{(m+n-r)}}{(m-r)!}.$$

Combining this with the result of (45) and (41), we see that

$$\begin{aligned}
\operatorname{cov}(S_1, K_{mn}) &= \mathbb{E}S_1 K_{mn} - \mathbb{E}S_1 \mathbb{E}K_{mn} \\
&= \operatorname{var} S_1 - \frac{\theta m!}{\theta_{(m+n)}} \sum_{r=1}^{m} \frac{1}{r} \frac{\theta_{(m+n-r)}}{(m-r)!} \left(1 + \sum_{i=0}^{m-r-1} \frac{\theta}{\theta + i} - \mathbb{E}S_1 \right) \\
&= \operatorname{var} S_1 - \frac{\theta m!}{\theta_{(m+n)}} \sum_{r=1}^{m} \frac{1}{r} \frac{\theta_{(m+n-r)}}{(m-r)!} \left(1 - \sum_{i=m-r}^{m-1} \frac{\theta}{\theta + i} \right),
\end{aligned}$$

which completes the proof. □

Remark. It is true, though it seems difficult to establish directly, that (9) is a symmetric function of m and n. See Lemma 2 below.

Remark. Applying (9) with $m = n$, we obtain $\mathbb{E}V_{2,n} = \operatorname{var} S_1 - \operatorname{cov}(S_1, S_2)$, which reduces to the result in (6) for $p = 2$.

3.3. Asymptotics for the covariance

In the next lemma, we examine the asymptotic behavior of the covariance in (9) in the regime in which $m, n \to \infty$ such that

$$n/m \to \beta \in (0, \infty);$$

its proof is given in the Appendix.

Lemma 1.

$$\operatorname{cov}(S_1(m), S_2(n)) = \operatorname{var} S_1(m) + \theta \log\left(\frac{n}{m+n} \right) + \frac{\theta^2}{n} + \frac{m}{m+n} \cdot \frac{\theta(\theta - \frac{1}{2})}{n} + O(m^{-2}),$$

where $\operatorname{var} S_1(m)$ *is given in (7) with n there replaced by m.*

Remark. The last lemma indicates that

$$\operatorname{cov}(S_1(m), S_2(n)) \approx \operatorname{var} S_1(m) + \theta \log\left(\frac{\beta}{1+\beta} \right) + \frac{1}{\beta m} \left(\frac{\theta(\theta - \frac{1}{2})}{1+\beta} + \theta^2 \right).$$

4. Multiple samples

In this section we exploit the exchangeable random partition feature of the Ewens Sampling Formula to establish the behavior of multiple samples, as opposed to the two considered above. Assume, then, that we take p samples sequentially, of sizes $n_1, n_2, \ldots, n_p$; the case treated above corresponds to $p = 2, n_1 = m, n_2 = n$. In particular, we will see that when $n_1 = n_2 = \cdots = n_p = n$ the sample counts $S_1, S_2, \ldots, S_p$ are exchangeable, and therefore from Lemma 2

$$\begin{aligned}\mathbb{E}\left\{\frac{1}{p-1}\sum_{j=1}^{p}(S_j-\bar{S})^2\right\} &= \frac{1}{p(p-1)}\sum_{i<j}\mathbb{E}(S_i-S_j)^2\\ &= \frac{1}{2}\,\mathbb{E}(S_1-S_2)^2\\ &= \sum_{r=n}^{2n-1}\frac{\theta}{\theta+r}-\frac{\theta n!}{\theta_{(2n)}}\sum_{r=1}^{n}\frac{1}{r}\frac{\theta_{(2n-r)}}{(n-r)!}\sum_{i=n-r}^{n-1}\frac{\theta}{\theta+i},\end{aligned}$$

as claimed in (6).

We write $n_+ = n_1 + \cdots + n_p$ for the total number of specimens sampled, and we label the n_+ specimens $1, 2, \ldots, n_+$, and set $n_0 = 0$. We generate a random partition X of specimens into species by running the sampling model for these n_+ specimens. For $i = 1, \cdots, p$, as before let S_i be the number of species identified in the ith sample.

Now suppose we rearrange the order in which these n_+ specimens are sampled. Let σ be a permutation on $[n_+] := \{1, 2, \ldots, n_+\}$, and consider the new sampling scheme where the order of sampling is given by $\sigma(1), \sigma(2), \cdots, \sigma(n_+)$. In other words, at time k, the specimen labeled $\sigma(k)$ is sampled and is either of a new species, or an existing species, chosen according to the existing species counts. Denote by X^σ the random partition generated by the sampling process with the rearranged sampling order σ.

For a subset $A = \{\ell_1, \cdots, \ell_k\} \subseteq [n_+]$, let $\sigma A = \{\sigma(\ell_1), \cdots, \sigma(\ell_k)\}$, and also for a partition $\pi = \{R_1, \cdots, R_b\}$ of $[n_+]$ define the partition $\sigma\pi := \{\sigma R_1, \cdots, \sigma R_b\}$. It is clear from the definition that for any partition π of $[n_+]$,

$$\mathbb{P}(X^\sigma = \pi) = \mathbb{P}(X = \sigma^{-1}\pi) = \mathbb{P}(\sigma X = \pi) = \mathbb{P}(X = \pi), \tag{10}$$

the last equality coming from the exchangeability of X; see Aldous [1, pp. 85, 92], Pitman [25, p. 56].

Indeed, from exchangeability, it is clear that rearranging the order of the samples does not effect the correlation and the joint distribution of each pair of samples. To be more precise, for $1 \le i \ne j \le p$, consider a permutation σ_{ij} that rearranges the order in which the specimens are sampled in such a way that the specimens in sample i, in order of their labels, are sampled first, then specimens of sample j, in order of their labels, are sampled next, and after these two samples are taken, the other specimens are sampled in order of their labels. From (10), $X \sim X^{\sigma_{ij}} \sim X^{\sigma_{ji}}$. In particular, the joint distribution of the number of species found in specimens $n_{i-1} + 1, \cdots, n_i$ and the number of species found in specimens $n_{j-1} + 1, \cdots, n_j$ does not depend on the order in which the specimens are sampled. The following lemma is an immediate consequence of exchangeability and symmetry.

Lemma 2. *For $1 \le i < j \le p$, $\mathcal{L}(S_i(n_i), S_j(n_j)) = \mathcal{L}(S_1(n_i), S_2(n_j))$. In particular*

(i) $\operatorname{cov}(S_i(n_i), S_j(n_j)) = \operatorname{cov}(S_1(n_i), S_2(n_j)) = \operatorname{cov}(S_1(n_j), S_2(n_i))$,
(ii) $\mathbb{E}(S_i(n_i) - S_j(n_j))^2 = \mathbb{E}(S_1(n_i) - S_2(n_j))^2 = \mathbb{E}(S_1(n_j) - S_2(n_i))^2$.

Remark. Lemma 2 indicates that the covariance formula in (9) given in Theorem 1 is, indeed, symmetric with respect to m and n. Also, similar lines of argument to those indicated before Lemma 2 imply that for $n_i = n$ for $i \in \mathbb{N}$, $(S_1, S_2, \cdots)$ is exchangeable.

The following symmetries can also be deduced immediately from exchangeability. Let $T'_{mn} := S_1(m) - K_{mn}$ be the number of species in the first sample of specimens $1, \ldots, m$ who are not present in the second sample. Note that T_{nm} and K_{nm} are defined by switching the roles of m and n in the original definitions such that the first and the second samples include n and m specimens, respectively. From exchangeability, $T'_{mn} \sim T_{nm}$ and $K_{mn} \sim K_{nm}$ in distribution. This can be established by letting the specimens of the second sample $m+1, \cdots, m+n$ be sampled first and the specimens of the first sample $1, \cdots, m$ next. As a result, $\mathbb{E}K_{mn} = \mathbb{E}K_{nm}$, and hence the formulae (8) and (40) are symmetric with respect to m and n, a fact that is also clear from the last equality in (41).

5. Asymptotics for the sample variance

We note that the value of

$$\mathbb{E}\left\{\frac{1}{p-1}\sum_{j=1}^{p}(S_j - \bar{S})^2\right\} = \frac{1}{p(p-1)}\sum_{i<j}\mathbb{E}(S_i - S_j)^2 \tag{11}$$

for arbitrary sample sizes $n_1, n_2, \ldots, n_p$ may be found from Theorem 1 by substitution of the appropriate sample sizes. Here, we record the asymptotic behavior in the case $n_i = lq_i$ for $i = 1, \ldots, p$ where $0 < q_i < 1$ and $q_1 + \cdots + q_p = 1$, as $l \to \infty$. Using Lemma 1 we see that for $i < j$,

$$\operatorname{cov}(S_i, S_j) \sim \theta \log(l) + \theta \log(q_i) + \theta \log(q_j/(q_i + q_j))$$

and hence

$$\begin{aligned}
\mathbb{E}(S_i - S_j)^2 &= \operatorname{var} S_i + \operatorname{var} S_j - 2\operatorname{cov}(S_i, S_j) + (\mathbb{E}S_i - \mathbb{E}S_j)^2 \\
&\sim -\theta \log\left(\frac{q_i}{q_i + q_j}\right) - \theta \log\left(\frac{q_j}{q_i + q_j}\right) + \theta^2(\log(q_i/q_j))^2 \\
&= \theta^2(\log(q_i/q_j))^2 + \theta \log\left(\frac{(q_i + q_j)^2}{q_i q_j}\right).
\end{aligned}$$

Substituting this into (11), we see that as $l \to \infty$,

$$\frac{1}{p(p-1)}\sum_{i<j}\mathbb{E}(S_i - S_j)^2 \sim \frac{1}{p(p-1)}\sum_{i<j}\left\{\theta^2(\log(q_i/q_j))^2 + \theta \log\left(\frac{(q_i + q_j)^2}{q_i q_j}\right)\right\}.$$

This formula was found by [5] by a Poisson process argument, without recourse to the pre-limiting formula we obtained here.

In the sequel, we focus on a more detailed analysis of the sample variance in the case of equal sample sizes, $n_i = n$, $i \in \mathbb{N}$, where as already noticed, the sequence $S_1, S_2, \cdots$ is exchangeable. We have seen so far that the expected value of the sample variance of $S_1, S_2, \ldots, S_p$, denoted by

$$V_{p,n} = \frac{1}{2p(p-1)}\sum_{i,j=1}^{p}\left(S_i - S_j\right)^2,$$

is approximately $\theta \log 2$ when n is large. In this section, we study the asymptotic behavior of $V_{p,n}$ when n is fixed and p tends to infinity (Section 5.1), when p is fixed and n tends to infinity (Section 5.2), and when n and $p = p_n$ both tend to infinity under some asymptotic regime (Section 5.3). Using the Poisson approximation provided in [5], for $i \neq j \in \mathbb{N}$, we construct some correlated random variables $\pi_{ij} \sim \mathrm{Po}(\theta \log 2)$, arising as the weak limit of the number of species identified in the ith sample that are absent in the jth sample, and show that, for fixed p, the sample variance converges in distribution to

$$\Phi_p := \frac{1}{2p(p-1)} \sum_{1 \leq i \neq j \leq p} (\pi_{ij} - \pi_{ji})^2.$$

In addition to providing a useful analysis of this Poisson system, under certain asymptotic regimes, when there exist b_n, for $n \in \mathbb{N}$, such that

$$\begin{aligned} p_n, b_n &\to \infty, \\ b_n/n &\to 0, \\ p_n^2 e^{-\frac{b_n}{p_n}}/b_n &\to 0, \end{aligned} \tag{12}$$

as $n \to \infty$, we prove that $V_{p,n}$ converges in distribution to a random variable with mean $\theta \log 2$ and variance $\theta \log(9/8)$. To establish this, let $\Psi_n = \mathbb{E}[V_{2,n} \mid \mathcal{F}_\infty^{(n)}]$ and $\Phi := \mathbb{E}[\Phi_2 \mid \mathcal{F}_\infty]$, where $\mathcal{F}_\infty^{(n)} = \cap_{p \geq 2} \sigma(V_{i,n} \ : \ i \geq p)$ and $\mathcal{F}_\infty = \cap_{p \geq 2} \sigma(\Phi_i \ : \ i \geq p)$. The main goal of this section is to prove the following result.

Theorem 2. *Under conditions (12),*

$$V_{p_n,n} \Rightarrow \Phi, \ \ n \to \infty.$$

Moreover, we have the following commutative convergence diagram:

$$\begin{array}{ccc} V_{p,n} & \xrightarrow{a.s., L^1} & \Psi_n \\ \Downarrow & \searrow & \Downarrow \\ \Phi_p & \xrightarrow[a.s., L^1]{} & \Phi \end{array}$$

as $n, p \to \infty$ appropriately.

The distribution of Φ is determined by its moments which can be computed from Lemma 4 and the definition of π_{ij} in (22) and (23). The expected value and the variance of Φ is computed in Proposition 2. To prove Theorem 2 (Section 5.3), we make use of some basic tools for exchangeable random variables and the Poisson approximation provided in [5].

5.1. Limit of sample variance when n is fixed and $p \to \infty$

Consider a map $g : \mathbb{R} \to \mathbb{R}$ s.t. $|g(S_1)|$ is integrable. From de Finetti's theorem for the exchangeable sequence $\Xi_n = (S_i)_{i \geq 1}$ ([18, Theorem 11.10], [1, Theorem 3.1]), we have as $p \to \infty$,

$$p^{-1} \sum_{i=1}^{p} g(S_i) \to \bar{g}(\nu_n) := \int_{\mathbb{R}_+} g(x) \nu_n(dx), \ \ \text{a.s.}, \tag{13}$$

where ν_n is the random measure on $\mathbb{R}$, a.s. uniquely determined by $\mathbb{P}(\mathfrak{S}^{(\mathbf{n})} \in \cdot \mid \nu_n) = \nu_n^{\otimes\mathbb{N}}(\cdot)$. Equivalently, from the Glivenko-Cantelli Theorem [18, Proposition 4.24], [1, p. 15], ν_n may be defined as the a.s. limit of the random empirical measures $p^{-1}\sum_{i=1}^{p}\delta_{S_i}$, in the weak topology of measures. Furthermore for $\ell \in \mathbb{N}$, $i = 1, \cdots, \ell$ and $g_i : \mathbb{R} \to \mathbb{R}$, $|\prod_{i=1}^{\ell} g_i(S_i)|$ integrable, we have

$$\begin{aligned} \mathbb{E}[\prod_{i=1}^{\ell} g_i(S_i) \mid \nu_n] &= \prod_{i=1}^{\ell} \bar{g}_i(\nu_n) \quad \text{a.s.}, \\ \mathbb{E}[\prod_{i=1}^{\ell} g_i(S_i)] &= \mathbb{E}[\prod_{i=1}^{\ell} \bar{g}_i(\nu_n)], \quad \text{(cf. [1, (2.22)])}. \end{aligned} \tag{14}$$

Letting $m(\nu_n) := \int_{\mathbb{R}_+} x\ \nu_n(dx)$, from (14), we have $\mathbb{E}(m(\nu_n)) = \mathbb{E}S_1$ and

$$\operatorname{var} m(\nu_n) = \mathbb{E}S_1S_2 - (\mathbb{E}S_1)^2 = \operatorname{cov}(S_1, S_2).$$

Applying $g(x) = x^r$ in (13), for $r = 1, 2$, and setting $\bar{S}_p = \bar{S}_p(n) := p^{-1}\sum_{i=1}^{p} S_i$, it follows that

$$V_{p,n} = \frac{1}{p-1}\left(\sum_{i=1}^{p} S_i^2 - p\bar{S}_p^2\right) \longrightarrow \widetilde{\Psi}_n := \int_{\mathbb{R}_+} x^2\ \nu_n(dx) - \left(\int_{\mathbb{R}_+} x\ \nu_n(dx)\right)^2 \quad \text{a.s.}, \tag{15}$$

as $p \to \infty$. From (14), we have

$$\mathbb{E}\widetilde{\Psi}_n = \mathbb{E}S_1^2 - \mathbb{E}S_1S_2 = \operatorname{var} S_1 - \operatorname{cov}(S_1, S_2),$$

which is computed in Theorem 1. Applying (14) again gives

$$\mathbb{E}\widetilde{\Psi}_n^2 = \mathbb{E}(S_1S_2)^2 + \mathbb{E}S_1S_2S_3S_4 - 2\mathbb{E}S_1^2S_2S_3, \tag{16}$$

and from (16), or directly from the pre-limit, we get

$$\operatorname{var}(\widetilde{\Psi}_n) = \lim_{p\to\infty} \operatorname{var}(V_{p,n}) = \operatorname{cov}\left(\frac{1}{2}(S_1 - S_2)^2, \frac{1}{2}(S_3 - S_4)^2\right). \tag{17}$$

The right hand side of (17) can be calculated using the frequency spectrum function method introduced in [11]. We omit the details here, as the calculation is somewhat involved.

The limit in (15) can also be obtained from the so-called weak exchangeability property of $\mathcal{S} = (\mathcal{S}_{ij})_{i\neq j\in\mathbb{N}}$, where $\mathcal{S}_{ij} = \mathcal{S}_{ij}(n) = (S_i - S_j)^2/2$. Say a random array $(X_{i,j})_{i\neq j\in\mathbb{N}}$ is weakly exchangeable if $X_{i,j}$ is symmetric with respect to i and j, for any $i, j \in \mathbb{N}$, and if for any finite permutation σ in the infinite symmetric group

$$(X_{i,j})_{i\neq j\in\mathbb{N}} \sim (X_{\sigma(i),\sigma(j)})_{i\neq j\in\mathbb{N}},$$

in distribution. It is clear from the definition that $\mathcal{S}$ is weakly exchangeable. Letting $\mathcal{F}_r^{(n)} := \sigma(V_{i,n} : i \geq r)$ and recalling $\mathcal{F}_\infty^{(n)} = \cap_{r\geq 2}\mathcal{F}_r^{(n)}$, [9] observed that $V_{p,n} = \mathbb{E}(\mathcal{S}_{1,2} \mid \mathcal{F}_p^{(n)})$ and hence it is a reversed martingale with respect to filtration $\mathcal{F}_p^{(n)}$, $p \geq 2$. Then from [9, Theorem 3], we have

$$V_{p,n} \xrightarrow{a.s., L^1} \Psi_n = \mathbb{E}(\mathcal{S}_{1,2} \mid \mathcal{F}_\infty^{(n)}), \tag{18}$$

and therefore $\Psi_n = \widetilde{\Psi}_n$, a.s.. See [9, Theorem 4] for a central limit theorem. The proof of the following lemma is given in Section A.3 of the Appendix.

Lemma 3. *For any $k \in \mathbb{N}$,*

$$\mathbb{E}[\Psi_n^k] = \mathbb{E}[\prod_{i=1}^{k} \mathcal{S}_{2i-1,2i}]. \tag{19}$$

Remark. The null distribution of $V_{p,n}$ for small n and p was used in [5] as a test of homogeneity of the different samples. When p is large, the necessary simulation to find the required distribution is often impractical. We can, however, effectively simulate the moments in (19) for small values of k and then use these to approximate the tails of the distribution of Ψ_n (see [22,23] for example). A plug-in estimator might be used for the unknown parameter θ.

5.2. Poisson approximation for large n

For $p \in \mathbb{N}$ and $\varnothing \neq A \subsetneq [p]$, let $\bar{T}_A^{(p)} = \bar{T}_A^{(n,p)}$ be the number of those species who are present in every sample $i \in A$ and absent in every sample $j \in [p] \setminus A$. We denote by $\mathcal{T}_{ij} = \mathcal{T}_{ij}^{(n)}$ the number of all those species who are present in sample i, but absent in sample j, that is $\mathcal{T}_{12}^{(n)} = T_{nn}$, and from exchangeability $\mathcal{T}_{ij}^{(n)} \sim T_{nn}$ in distribution, for any $i \neq j$. The definition of $\mathcal{T}_{ij}$ does not depend on the number of samples p, and we can, in fact, consider an infinite number of samples $S_1, S_2, \ldots$. In other words, for any $p \in \mathbb{N}$,

$$\mathcal{T}_{ij} = \sum_{A \subsetneq [p]: i \in A, j \notin A} \bar{T}_A^{(p)}. \tag{20}$$

As $S_i - S_j = \mathcal{T}_{ij} - \mathcal{T}_{ji}$, to study the asymptotic behavior of the sample variance, it is most convenient to study that of $(\bar{T}_A^{(p)} : \varnothing \neq A \subsetneq [p])$. We find the expected value of $\bar{T}_A$. Again from exchangeability, the distribution of $\bar{T}_A$ depends on the cardinality of A, not A itself. By the inclusion-exclusion principle, for $|A| = r$, we have

$$\begin{aligned}
\mathbb{E}[\bar{T}_A^{(p)}] &= \sum_{\ell=0}^{r-1} (-1)^\ell \binom{r}{\ell} \mathbb{E} T_{n(p-r+\ell),n(r-\ell)} \\
&= \sum_{\ell=0}^{r-1} (-1)^\ell \binom{r}{\ell} \sum_{i=n(p-r+\ell)+1}^{n_+} \frac{\theta}{\theta + i - 1} \\
&= \sum_{\ell=0}^{r-1} q_{r,\ell} \sum_{i=n(p-r+\ell)+1}^{n(p-r+\ell+1)} \frac{\theta}{\theta + i - 1},
\end{aligned}$$

where, for $0 \leq \ell < r$

$$q_{r,\ell} := \sum_{j=0}^{\ell} (-1)^j \binom{r}{j}.$$

Thus

$$\tilde{\lambda}_r^{(p)} = \bar{\lambda}_A^{(p)} := \lim_{n\to\infty} \mathbb{E}[\bar{T}_A^{(n,p)}] = \theta \sum_{\ell=0}^{r-1} q_{r,\ell} \log\left(\frac{p-r+\ell+1}{p-r+\ell}\right)$$
$$= \theta \sum_{\ell=0}^{r} (-1)^{\ell+1} \binom{r}{\ell} \log(p-r+\ell),$$

since $q_{r,\ell-1} - q_{r,\ell} = (-1)^{\ell+1}\binom{r}{\ell}$ for $\ell = 1, \cdots, r-1$, $q_{r,0} = \binom{r}{0}$ and

$$q_{r,r-1} = \sum_{j=0}^{r-1} (-1)^j \binom{r}{j} = 0 - (-1)^r \binom{r}{r} = (-1)^{r+1} \binom{r}{r}.$$

For example, for $p = 4$ we have $\tilde{\lambda}_1^{(4)} = \theta \log(4/3)$, $\tilde{\lambda}_2^{(4)} = \theta \log(9/8)$ and $\tilde{\lambda}_3^{(4)} = \theta \log(32/27)$.

Consider independent Poisson random variables $\Pi_p := (\bar{\pi}_A^{(p)} : \varnothing \neq A \subsetneq [p])$ with $\bar{\pi}_A^{(p)} \sim Po(\bar{\lambda}_A^{(p)})$. For a given p, Barbour and Tavaré [5] showed that $\boldsymbol{T}_{n,p} := (\bar{T}_A^{(n,p)} : \varnothing \neq A \subsetneq [p])$ are asymptotically independent and

$$d_{TV}(\mathcal{L}(\boldsymbol{T}_{n,p}), \mathcal{L}(\Pi_p)) \leq \frac{bc_p}{np} + \frac{p^2(3\theta + 1/e)(1-1/p)^{b+1}}{b+1}, \tag{21}$$

where $4(b+1) \leq np$ and $c_p = \max(4, 4\theta(p+1+\theta)/3)$. Taking $b = b_n$ such that $b_n \to \infty$ and $b_n/n \to 0$, as $n \to \infty$, guarantees that

$$\boldsymbol{T}_{n,p} \Rightarrow \Pi_p, \quad n \to \infty.$$

We discuss the case $n, p \to \infty$ later. Inspired by (20), we define

$$\pi_{ij}^{(p)} := \sum_{A \subsetneq [p]: i \in A, j \notin A} \bar{\pi}_A^{(p)}. \tag{22}$$

It is not hard to see that the distributions of $\pi_{ij}^{(p)}$ do not depend on p (cf. Proposition 1). In fact, for any p

$$(\pi_{ij}^{(p)})_{i \neq j \leq p} \sim (\pi_{ij}^{(p+1)})_{i \neq j \leq p},$$

and thus there exists a projective limit sequence $(\pi_{ij})_{i \neq j \in \mathbb{N}}$ such that

$$(\pi_{ij})_{i \neq j \leq p} \sim (\pi_{ij}^{(p)})_{i \neq j \leq p}, \tag{23}$$

for any p. It is clear from (21) that $(\mathcal{T}_{ij}^{(n)})_{i \neq j \leq p}$ converges weakly to $(\pi_{ij})_{i \neq j \leq p}$ as $n \to \infty$, for any given $p \in \mathbb{N}$, and therefore

$$(\mathcal{T}_{ij}^{(n)})_{i \neq j \in \mathbb{N}} \Rightarrow (\pi_{ij})_{i \neq j \in \mathbb{N}}, \quad n \to \infty.$$

To ease the notation, for $A = \{i_1, \cdots, i_\ell\}$, we let $\bar{\pi}_{i_1 \cdots i_\ell}^{(p)} = \bar{\pi}_A^{(p)}$; e.g. $\bar{\pi}_{12}^{(4)}$ stands for $\bar{\pi}_{\{1,2\}}^{(4)}$.

Proposition 1. *For any different natural numbers i, j, k, l, we have $\pi_{ij} \sim Po(\theta \log(2))$, and*

i) $\operatorname{cov}(\pi_{ij}, \pi_{kl}) = \theta \log(9/8)$,

ii) $\operatorname{cov}(\pi_{ij},\pi_{ik}) = \theta\log(3/2)$,
iii) $\operatorname{cov}(\pi_{ij},\pi_{kj}) = \theta\log(4/3)$,
iv) $\operatorname{cov}(\pi_{ij},\pi_{jk}) = 0$.

Proof. We first show that, for any $p \in \mathbb{N}$, $\pi_{ij}^{(p)} \sim \pi_{ij}^{(p+1)}$. We need to show that $\lambda_{ij}^{(p)} = \lambda_{ij}^{(p+1)}$, where

$$\lambda_{ij}^{(p)} := \sum_{A\subsetneq[p]:i\in A, j\notin A} \bar{\lambda}_A^{(p)} = \sum_{r=0}^{p-2}\binom{p-2}{r}\tilde{\lambda}_{r+1}^{(p)}$$

for $i \neq j \in [p]$. But for $|A| = r < p$ this follows from

$$\begin{aligned}\bar{\lambda}_A^{(p+1)} + \bar{\lambda}_{A\cup\{p+1\}}^{(p+1)} =& \theta\sum_{\ell=0}^{r+1}(-1)^{\ell+1}\binom{r+1}{\ell}\log(p-r+\ell)\\ &+\theta\sum_{\ell=0}^{r}(-1)^{\ell+1}\binom{r}{\ell}\log(p+1-r+\ell)\\ =&\theta\sum_{\ell=0}^{r+1}(-1)^{\ell+1}\left(\binom{r+1}{\ell}-\binom{r}{\ell-1}\right)\log(p-r+\ell) = \bar{\lambda}_A^{(p)},\end{aligned}$$

since $\binom{r+1}{\ell} - \binom{r}{\ell-1} = \binom{r}{\ell}$. Therefore, for any $p \in \mathbb{N}$, $\lambda_{ij}^{(p)} = \lambda_{12}^{(2)} = \theta\log 2$, and this yields $\pi_{ij}^{(p)} \sim \mathrm{Po}(\theta\log 2)$. So $\pi_{ij} \sim \mathrm{Po}(\theta\log 2)$. To prove (i), note that from symmetry and the independence of $\bar{\pi}_A^{(p)}$, for $A \subsetneq [p]$, we have

$$\begin{aligned}\operatorname{cov}(\pi_{ij},\pi_{kl}) =& \operatorname{cov}(\pi_{12},\pi_{34})\\ =&\operatorname{cov}(\bar{\pi}_1^{(4)}+\bar{\pi}_{13}^{(4)}+\bar{\pi}_{14}^{(4)}+\bar{\pi}_{134}^{(4)},\bar{\pi}_3^{(4)}+\bar{\pi}_{13}^{(4)}+\bar{\pi}_{23}^{(4)}+\bar{\pi}_{123}^{(4)})\\ =&\operatorname{var}(\bar{\pi}_{13}^{(4)}) = \theta\log(9/8).\end{aligned}$$

Similarly, we have

$$\operatorname{cov}(\pi_{ij},\pi_{ik}) = \operatorname{cov}(\pi_{12},\pi_{13}) = \operatorname{cov}(\bar{\pi}_1^{(3)}+\bar{\pi}_{13}^{(3)},\bar{\pi}_1^{(3)}+\bar{\pi}_{12}^{(3)}) = \operatorname{var}(\bar{\pi}_1^{(3)}) = \theta\log(3/2),$$

$$\operatorname{cov}(\pi_{ij},\pi_{kj}) = \operatorname{cov}(\pi_{12},\pi_{32}) = \operatorname{cov}(\bar{\pi}_1^{(3)}+\bar{\pi}_{13}^{(3)},\bar{\pi}_3^{(3)}+\bar{\pi}_{13}^{(3)}) = \operatorname{var}(\bar{\pi}_{13}^{(3)}) = \theta\log(4/3),$$

and

$$\operatorname{cov}(\pi_{ij},\pi_{jk}) = \operatorname{cov}(\pi_{12},\pi_{23}) = \operatorname{cov}(\bar{\pi}_1^{(3)}+\bar{\pi}_{13}^{(3)},\bar{\pi}_2^{(3)}+\bar{\pi}_{12}^{(3)}) = 0.$$ □

Let $\varphi_{ij} := \pi_{ij} - \pi_{ji}, \phi_{ij} := \varphi_{ij}^2/2$, and recall that

$$\Phi_p := \frac{1}{p(p-1)}\sum_{1\le i\neq j\le p}\phi_{ij}.$$

Note that, similarly to $\mathcal{S}_{ij}$ in the previous section, ϕ_{ij} are weakly exchangeable, and from [9, Theorem 3] and the above discussion

$$\Phi_p \xrightarrow{a.s., L^1} \Phi = \mathbb{E}[\phi_{1,2} \mid \mathcal{F}_\infty],\ p\to\infty, \tag{24}$$

where as defined before $\mathcal{F}_\infty = \cap_{r\geq 2}\sigma(\Phi_i : i \geq r)$. From [9, Theorem 4] we find

$$\sqrt{p}(\Phi_p - \Phi) \Rightarrow \mathcal{N}^*, \ p \to \infty,$$

where the law of $\mathcal{N}^*$ is the mixture of standard normal distributions with characteristic function

$$t \mapsto \mathbb{E}[\exp\{-2t^2 \mathrm{cov}(\phi_{1,2}, \phi_{1,3} \mid \mathcal{F}_\infty)\}].$$

The law of φ_{ij}, which is the difference of two independent Poisson-distributed random variables, is called the Skellam distribution [16,27]. From [16],

$$\mathbb{P}(\varphi_{ij} = r) = e^{-2\theta \log 2} \mathcal{I}_r(2\theta \log 2),$$

where $\mathcal{I}_r$ is the modified Bessel function of the first kind defined by

$$\mathcal{I}_r(x) = \left(\frac{x}{2}\right)^r \sum_{j=0}^{\infty} \frac{x^{2j}}{4^j j!(r+j)!}.$$

The moment generating function of φ_{ij} is given by

$$\mathcal{M}_{ij}(t) := \mathbb{E}[e^{t\varphi_{ij}}] = \exp\{2\theta \log(2)(\cosh(t) - 1)\}, \ \ (\text{cf. } [16]).$$

We finish this section with the following results, whose proofs are given in the Appendix.

Lemma 4. *The distribution of* Φ *is determined by its moments, and for* $k \in \mathbb{N}$

$$\mathbb{E}[\Phi^k] = \frac{1}{2^k}\mathbb{E}\left[\prod_{i=1}^{k}(\pi_{2i-1,2i} - \pi_{2i,2i-1})^2\right]. \tag{25}$$

Proposition 2. $\mathbb{E}\Phi = \theta \log 2$, *and* $\mathrm{var}(\Phi) = \lim_{p\to\infty} \mathrm{var}(\Phi_p) = \mathrm{cov}(\phi_{1,2}, \phi_{3,4}) = \theta \log(9/8)$.

Remark. Following the remark after (19), the moments in (25) may be simulated and used to approximate the tails of the distribution of Φ.

5.3. Limit diagram for the sample variance

Before giving the proof of Theorem 2, we analyze the upper bound (21) for the total variation distance a bit more. Consider the case that $b = b_n, p = p_n \to \infty$ as n tends to infinity. For large n, the upper bound in (21) can be approximated by

$$\frac{4\theta b_n}{3n} + \frac{(3\theta + e^{-1})p_n^2(1 - \frac{1}{p_n})^{b_n}}{b_n}.$$

In order for this expression to converge to 0, it suffices to have $b = b_n$ such that (12) holds as $n \to \infty$. But it is not hard to see that (12) holds if $p_n = O(n^{1-\varepsilon})$, for some $0 < \varepsilon < 1$. This can be handled by letting $b_n = n^{1-\varepsilon'}$ for $0 < \varepsilon' < \varepsilon$. On the other hand, although $p_n = o(b_n)$ and $p_n = o(n)$ are necessary conditions for (12), neither of them is sufficient. To see this, let $b_n/p_n = \log\log n$. Then

$$\frac{p_n^2 e^{-\frac{b_n}{p_n}}}{b_n} = \frac{p_n e^{-\log\log n}}{\log\log n} = \frac{p_n}{\log n \log\log n}, \tag{26}$$

tending to ∞, as $n \to \infty$, for an appropriate choice of p_n. In particular, for $p_n = n/\log n$ and $b_n = n \log\log n/\log n$, for any $0 < \varepsilon < 1$, we get $n^{1-\varepsilon} \ll p_n \ll b_n \ll n$, but in this case, the right of (26) is $n/\{(\log n)^2 \log\log n\} \to \infty$, as $n \to \infty$.

Proof of Theorem 2. The a.s. and L^1 convergence is discussed in (15), (18), and (24). Letting $b = b_n$ in (21) such that $b_n \to \infty$ and $b_n/n \to 0$, as $n \to \infty$, we get $\boldsymbol{T}_{n,p} \Rightarrow \Pi_p$, and hence from the continuous mapping theorem, $V_{p,n} \Rightarrow \Phi_p$, as $n \to \infty$. On the other hand, it follows from (19) and Lemma 4 that

$$\begin{aligned}\mathbb{E}[\Psi_n^k] =& \mathbb{E}\left[\prod_{i=1}^{k} \frac{1}{2}(S_{2i-1} - S_{2i})^2\right] \\ =& \mathbb{E}\left[\prod_{i=1}^{k} \frac{1}{2}(\mathcal{T}^{(n)}_{2i-1,2i} - \mathcal{T}^{(n)}_{2i,2i-1})^2\right] \\ \to& \mathbb{E}\left[\prod_{i=1}^{k} \phi_{2i-1,2i}\right] = \mathbb{E}[\Phi^k],\end{aligned}$$

for any $k \in \mathbb{Z}_+$, and hence we conclude $\Psi_n \Rightarrow \Phi$, as $n \to \infty$. Finally, taking b_n, p_n such that the conditions (12) hold, the diagonal weak convergence follows from $d_{TV}(V_{p_n,n}, \Phi_{p_n}) \to 0$ and $\Phi_{p_n} \Rightarrow \Phi$, as $n \to \infty$. □

6. Species present in all samples

Let p and n be the number of samples and the number of specimens for each sample, respectively. We study the number $\mathcal{K}_{n,p}$ of species that are present in all these p samples. To find the limit of $S_i(n) - \mathcal{K}_{n,p}$, we explore the behavior of

$$\pi_i^{(p)} = \sum_{i \in A \subsetneq [p]} \bar{\pi}_A,$$

for large n. From independence of the Poisson r.v. $\bar{\pi}_A$, $\pi_i^{(p)} = \mathrm{Po}(\theta\alpha_p)$, where

$$\alpha_p = \sum_{\ell=0}^{p-2} \binom{p-1}{\ell} \tilde{\lambda}^{(p)}_{\ell+1},$$

and in fact

$$\alpha_\infty := \lim_{p\to\infty} \alpha_p = \sum_{p=2}^{\infty} \tilde{\lambda}^{(p)}_{p-1}.$$

The limit of $\pi_i^{(p)}$, namely π_i, exists and is $\mathrm{Po}(\theta\alpha_\infty)$-distributed if and only if $\alpha_\infty < \infty$. But determining α_∞ from the above sums does not seem easy. For fixed n and p, by the inclusion-exclusion principle, we calculate instead

$$\mathbb{E}[\mathcal{K}_{n,p}] = \mathbb{E}[S_1(n)] - \sum_{\ell=1}^{p-1} (-1)^{\ell+1} \binom{p-1}{\ell} \mathbb{E}[T_{\ell n,n}],$$

so that

$$\theta\alpha_p = \lim_{n\to\infty}\sum_{\ell=1}^{p-1}(-1)^{\ell+1}\binom{p-1}{\ell}\mathbb{E}[T_{\ell n,n}]$$

$$=\theta\sum_{\ell=1}^{p-1}(-1)^{\ell+1}\binom{p-1}{\ell}\log\left(1+\frac{1}{\ell}\right).$$

The next result, whose proof appears in the appendix, shows that $\alpha_p \to \infty$ as $p \to \infty$.

Proposition 3. *For any $p > 1$,*

$$\alpha_p = \log\log p + \gamma + \frac{\gamma}{\log p} + \frac{1}{p\log p} + O(\log^{-2} p),$$

where $\gamma = 0.57721\cdots$ is Euler's constant.

Choose p_n so that conditions (12) hold. It then follows from (21) that, for any $i \in [p_n]$

$$d_{TV}(S_i(n) - \mathcal{K}_{n,p_n}, \mathcal{X}_n) \to 0,\ n \to \infty, \tag{27}$$

where $\mathcal{X}_n = \mathcal{X}_{n,p_n} \sim \mathrm{Po}(\alpha_{p_n})$. Letting $\mathcal{A}_{n,p_n}$ be the number of species who are absent in at least one of the p_n samples, (27) indicates that for large n, $\mathcal{A}_{n,p_n}$ can be approximated by a Poisson random variable with parameter

$$\beta_n = \beta_{n,p_n} := \sum_{i=0}^{np_n-1}\frac{\theta}{\theta+i} - \left(\sum_{i=0}^{n-1}\frac{\theta}{\theta+i} - \alpha_{p_n}\right) \sim\theta\{\log(np_n) - (\log(n) - \alpha_{p_n})\}$$

$$\sim\theta\{\log p_n + \log\log p_n + \gamma\},$$

while $\mathcal{K}_{n,p_n}$ can be approximated by a Poisson random variable of parameter $\theta(\log(n) - \alpha_{p_n})$, which, roughly speaking, means the majority of the species in p_n samples are present in all samples as n grows. The next theorem follows immediately.

Theorem 3. *Suppose n and p_n satisfy in conditions (12). Then as $n \to \infty$,*

$$\frac{\mathcal{K}_{n,p_n} - \theta\log n}{\sqrt{\theta\log n}} \Longrightarrow \mathcal{N},$$

$$\frac{\mathcal{A}_{n,p_n} - \theta\log p_n}{\sqrt{\theta\log p_n}} \Longrightarrow \mathcal{N},$$

$$\frac{S_i(n) - \mathcal{K}_{n,p_n} - \theta\log\log p_n}{\sqrt{\theta\log\log p_n}} \Longrightarrow \mathcal{N},$$

where $\mathcal{N} \sim N(0,1)$.

Proof. It is well-known that (see [4])

$$\frac{S_1(n) - \theta\log n}{\sqrt{\theta\log n}} \Rightarrow \mathcal{N}, n \to \infty.$$

From (12), we must have $p_n = o(n)$, so that $\log\log p_n/\sqrt{\log n} \to 0$, as $n \to \infty$. Therefore, from Chebyshev's inequality we get

$$\frac{\mathcal{X}_n}{\sqrt{\theta \log n}} \xrightarrow{p} 0,$$

and hence (27) and Theorem 25.4 in [6] imply the first limit. For the second and third limits, note that from (21),

$$d_{TV}(\mathcal{A}_{n,p_n}, \mathcal{Y}_n) \to 0;\ n \to \infty, \tag{28}$$

where $\mathcal{Y}_n \sim \mathrm{Po}(\beta_n)$. The second and third limits follow as a result of the total variation estimates (27) and (28), and weak convergence of $(\mathcal{X}_n - \theta \log\log p_n)/\sqrt{\theta \log\log p_n}$ and $(\mathcal{Y}_n - \theta \log p_n)/\sqrt{\theta \log p_n}$ to $\mathcal{N}$. □

7. Fisher's log-series distribution revisited

In this section we consider the behavior of the species newly discovered in a second sample of size n following a first sample of size m. We focus in particular on the asymptotic regime described earlier, in which

$$m \to \infty, n \to \infty, n/m \to \beta \in (0, \infty). \tag{29}$$

We show *inter alia* that, asymptotically, the distribution of the number of specimens that belong to each species not found among the first m specimens behaves like a sample from a sequence of independent and identically distributed log-series random variables. This result is made precise in Theorem 4 below.

First we identify the distribution of the number of specimens $m+1, \ldots, m+n$ that belong to species not found in specimens $1, 2, \ldots, m$. Defining V_{mn} to be the number of extra specimens who are of new species, excluding the T_{mn} which are the type specimen for their species, we have, for $0 \le t \le n, 0 \le v \le n-t, n = 1, 2, \ldots$

$$\mathbb{P}(T_{mn} = t, V_{mn} = v) = \frac{\theta^t}{(\theta + m)_{(n)}} m_{(n-t-v)} \binom{n}{t+v} \begin{bmatrix} t+v \\ t \end{bmatrix}. \tag{30}$$

This may be derived from [21, Eqn. (10)]; see also [26] for the connection to the three-parameter generalized Stirling numbers. We provide a simple, direct combinatorial proof. Consider specimens $m+1, \ldots, m+n$ arriving according to the sampling process. The probability of any given assignment of the n specimens to species, with t new species, is $\theta^t/(m+\theta)_{(n)}$. There are $\binom{n}{t+v}$ different sets of specimens who can be chosen to be of the new species, and $\left[\begin{smallmatrix} t+v \\ t \end{smallmatrix}\right]$ ways these specimens can be formed into the t identified species. Finally, there are $m_{(n-t-v)}$ ways in which the $n-t-v$ specimens who are of the species identified in the first m specimens can be allocated. Multiplying these together gives the result.

It follows that

$$\mathbb{P}(T_{mn} = t) = \frac{\theta^t}{(m+\theta)_{(n)}} \sum_{j=t}^{n} \binom{n}{j} m_{(n-j)} \begin{bmatrix} j \\ t \end{bmatrix}, \quad t = 0, 1, \ldots, n.$$

Letting $D_{mn} = T_{mn} + V_{mn}$ denote the number of specimens $m+1, \ldots, m+n$ who represent new species, it follows from (30) that the distribution of D_{mn} is hypergeometric: for $d = 0, 1, \ldots, n$,

$$\mathbb{P}(D_{mn} = d) = \binom{n}{d} \frac{m_{(n-d)}\, \theta_{(d)}}{(m+\theta)_{(n)}}$$

$$= \binom{\theta+d-1}{d}\binom{m+n-d-1}{n-d}\Big/\binom{m+n+\theta-1}{n} \tag{31}$$

The mean and variance of D_{mn} are given by

$$\mathbb{E}D_{mn} = \frac{n\theta}{m+\theta}, \qquad \operatorname{var} D_{mn} = \frac{\theta nm}{(\theta+m)^2}\,\frac{\theta+m+n}{\theta+m+1}.$$

It follows from (30) that, for $0 \le t \le d \le n$,

$$\mathbb{P}(T_{mn}=t, D_{mn}=d) = \frac{\theta^t}{(\theta+m)_{(n)}}\, m_{(n-d)}\binom{n}{d}\begin{bmatrix} d \\ t \end{bmatrix}, \tag{32}$$

and so

$$\mathbb{P}(T_{mn}=t \mid D_{mn}=d) = \frac{\theta^t \begin{bmatrix} d \\ t \end{bmatrix}}{\theta_{(d)}}, t = 1,2,\ldots,d.$$

This shows that, given there are d specimens identified as species not found in the first m specimens, the number of distinct new species has the same distribution as that arising from an ESF with sample size d. This is a consequence of exchangeability, and in fact the distribution of the number of new species represented once, twice, ... has the ESF(θ) distribution with sample size d; cf. [21]. This setting provides a natural example where, in the language of the introduction, the number of species (S) and the number of specimens (N) are both random.

Under the limiting regime in (29), the asymptotic joint distribution of (T_{mn}, D_{mn}) is given by

Lemma 5. *As $m \to \infty, n \to \infty$ and $n/m \to \beta$,*

$$(T_{mn}, D_{mn}) \Rightarrow (T_\beta, D_\beta)$$

where

$$\mathbb{P}(T_\beta = t, D_\beta = d) = \frac{\theta^t \begin{bmatrix} d \\ t \end{bmatrix}}{d!}\left(\frac{1}{1+\beta}\right)^\theta \left(\frac{\beta}{1+\beta}\right)^d, \quad 0 \le t \le d, d = 0,1,\ldots \tag{33}$$

The marginal distribution of D_β is negative binomial, with

$$\mathbb{P}(D_\beta = d) = \frac{\theta_{(d)}}{d!}\left(\frac{1}{1+\beta}\right)^\theta \left(\frac{\beta}{1+\beta}\right)^d, \quad d = 0,1,\ldots$$

and the marginal distribution of T_β is Poisson, with

$$\mathbb{P}(T_\beta = t) = e^{-\lambda}\frac{\lambda^t}{t!}, \quad t = 0,1,\ldots, \tag{34}$$

where

$$\lambda = \theta \log(1+\beta).$$

Proof. Note first that for fixed d,

$$\frac{n!}{(n-d)!}\,\frac{m_{(n-d)}}{(\theta+m)_{(n)}} = \frac{\Gamma(n+1)}{\Gamma(n+1-d)}\,\frac{\Gamma(m+n-d)}{\Gamma(m+n+\theta)}\,\frac{\Gamma(m+\theta)}{\Gamma(m)}$$

$$\sim n^{1-(1-d)}\, m^{\theta}\, (m+n)^{-d-\theta}$$
$$\to \left(\frac{1}{1+\beta}\right)^{\theta} \left(\frac{\beta}{1+\beta}\right)^{d}.$$

Substituting this into (32) shows that

$$\mathbb{P}(T_{mn}=t, D_{mn}=d) \to \mathbb{P}(T_\beta = t, D_\beta = d), \text{ given in (33).}$$

Summing (33) over $t = 0,\ldots,d$ gives

$$\mathbb{P}(D_\beta = d) = \frac{1}{d!}\left(\frac{1}{1+\beta}\right)^{\theta} \left(\frac{\beta}{1+\beta}\right)^{d} \sum_{t=0}^{d} \theta^t \begin{bmatrix} d \\ t \end{bmatrix}$$
$$= \frac{\theta_{(d)}}{d!}\left(\frac{1}{1+\beta}\right)^{\theta} \left(\frac{\beta}{1+\beta}\right)^{d},$$

while summing (33) over $d \geq t$ gives

$$\mathbb{P}(T_\beta = t) = \left(\frac{1}{1+\beta}\right)^{\theta} \theta^t \sum_{d\geq t} \frac{1}{d!} \begin{bmatrix} d \\ t \end{bmatrix} \left(\frac{\beta}{1+\beta}\right)^{d}$$
$$= \left(\frac{1}{1+\beta}\right)^{\theta} \theta^t \frac{\log^t(1+\beta)}{t!}$$
$$= e^{-\theta\log(1+\beta)} \frac{(\theta\log(1+\beta))^t}{t!}, \quad t = 0,1,\ldots,$$

completing the proof of the lemma. □

The mean and variance of D_β are given by

$$\mathbb{E}D_\beta = \beta\theta, \qquad \mathrm{var}D_\beta = \theta\beta(1+\beta).$$

We have seen that asymptotically the number of new species T_β identified in specimens $m+1,\ldots,m+n$ has a Poisson distribution with mean $\theta\log(1+\beta)$. Of course, each of these species has a number of representatives, so that the total number of specimens belonging to new species is a random sum of the form $D_\beta = X_1 + \cdots + X_{T_\beta}$. We ignore the trivial case in which $T_\beta = 0$, as then $D_\beta = 0$ as well. In this section we establish

Theorem 4.

(a) *Write $A_i(m,n)$ for the number of specimens belonging to the ith new species discovered among specimens $m+1,\ldots,m+n$. Then for $a_1 \geq 1, a_2 \geq 1, \ldots, a_t \geq 1, a_1 + \cdots + a_t = d, 1 \leq t \leq d \leq n$,*

$$\mathbb{P}(A_1(m,n) = a_1, \ldots, A_t(m,n) = a_t, T_{mn} = t, D_{mn} = d) =$$
$$\frac{\theta^t n!}{(\theta+m)_{(n)}} \frac{m_{(n-d)}}{(n-d)!} \frac{1}{a_t(a_t+a_{t-1})\cdots(a_t+\cdots+a_1)}.$$

(b) *As $m,n \to \infty, n/m \to \beta \in (0,\infty)$,*

$$\mathbb{P}(A_1(m,n) = a_1, \ldots, A_t(m,n) = a_t, T_{mn} = t, D_{mn} = d)$$

$$\to \mathbb{P}(A_1 = a_1, \ldots, A_t = a_t, T_\beta = t, D_\beta = d)$$
$$= \theta^t \left(\frac{\beta}{1+\beta}\right)^d \left(\frac{1}{1+\beta}\right)^\theta \frac{1}{a_t(a_t + a_{t-1}) \cdots (a_t + \cdots + a_1)}. \tag{35}$$

(c) *It follows that, for* $t \geq 1$,

$$\mathbb{P}(A_1 = a_1, \ldots, A_t = a_t, D_\beta = d \mid T_\beta = t) =$$
$$\left(\frac{\beta}{1+\beta}\right)^d \frac{t!}{\log^t(1+\beta)} \frac{1}{a_t(a_t + a_{t-1}) \cdots (a_t + \cdots + a_1)} \tag{36}$$

(d) *The distribution in (36) is the size-biassed law of* t *independent, identically distributed random variables* $X_1, \ldots, X_t$*, each having Fisher's log-series distribution*

$$\mathbb{P}(X_i = l) = \frac{1}{\log(1+\beta)} \left(\frac{\beta}{1+\beta}\right)^l \frac{1}{l}, \; l = 1, 2, \ldots.$$

Proof. (a) We have

$$\begin{aligned}
&\mathbb{P}(A_1(m,n) = a_1, \ldots, A_t(m,n) = a_t, T_{mn} = t, D_{mn} = d) \\
&= \mathbb{P}(A_1(m,n) = a_1, \ldots, A_t(m,n) = a_t, T_{mn} = t \mid D_{mn} = d) \mathbb{P}(D_{mn} = d) \\
&= \left\{ \frac{\theta^t}{\theta_{(d)}} \frac{d!}{a_t(a_t + a_{t-1}) \cdots (a_t + \cdots + a_1)} \right\} \frac{\theta_{(d)}}{d!} \frac{m_{(n-d)}}{(n-d)!} \frac{n!}{(\theta + m)_{(n)}} \\
&= \frac{\theta^t n!}{(\theta + m)_{(n)}} \frac{m_{(n-d)}}{(n-d)!} \frac{1}{a_t(a_t + a_{t-1}) \cdots (a_t + \cdots + a_1)},
\end{aligned}$$

the second equality coming from (31) and the distribution of the lengths of the ordered species counts in the ESF, given in Arratia, Barbour and Tavaré [3, Section 5.4].

(b) This follows using the same steps as the proof of Lemma 5.

(c) Divide (35) by $\mathbb{P}(T_\beta = t)$ given in (34) and simplify.

(d) Assume that $T_\beta = t \geq 1$, and fix $a_1 \geq 1, \ldots, a_t \geq 1$ and let $d = a_1 + \cdots + a_t$. Let $X_1, \ldots, X_t$ be i.i.d. log-series-distributed random variables. The probability that $X_1, X_2, \ldots, X_t$ result in observations $\{a_1, \ldots, a_t\}$ in some order is

$$t! \prod_{i=1}^{t} \frac{1}{\log(1+\beta)} \left(\frac{\beta}{1+\beta}\right)^{a_i} \frac{1}{a_i} = \frac{t!}{\log^t(1+\beta)} \left(\frac{\beta}{1+\beta}\right)^d \frac{1}{a_1 \cdots a_t},$$

so we can write (36) in the form

$$\mathbb{P}(A_1 = a_1, \ldots, A_t = a_t, D_\beta = d \mid T_\beta = t) =$$
$$\frac{t!}{\log^t(1+\beta)} \left(\frac{\beta}{1+\beta}\right)^d \frac{1}{a_1 a_2 \cdots a_t} \left\{ \frac{a_1 a_2 \cdots a_t}{a_t(a_t + a_{t-1}) \cdots (a_t + \cdots + a_1)} \right\}. \tag{37}$$

The term in $\{\}$ on the right of (37) may be written as

$$\frac{a_1}{d} \frac{a_2}{d - a_1} \frac{a_3}{(d - a_1 - a_2)} \cdots \frac{a_{t-1}}{d - a_1 - \cdots - a_{t-2}} \frac{a_t}{a_t},$$

the probability that a size-biassed sample results in choosing $A_1 = a_1, A_2 = a_2, \ldots, A_t = a_t$, as required.

Appendix

In the next section we collect a number of technical results that are used in the proof of Theorem 1.

A.1. Results for proving Theorem 1

We collect together the ingredients we need to compute the covariance of $S_1(m)$ and $S_2(n)$. Consider a species represented by a specimens in the first sample. The probability that no further specimens of this species are found in specimens $m+1, m+2, \ldots, m+n$ is

$$\prod_{l=0}^{n-1} \frac{\theta+m+l-a}{\theta+m+l} = \frac{(\theta+m-a)_{(n)}}{(\theta+m)_{(n)}}. \tag{38}$$

Suppose we are given $S_1(m)=k$, the first species detected having $A_1(m)=a_1$ specimens, the second $A_2(m)=a_2$, the kth having $A_k(m)=a_k$, where $a_1 \geq 1, \ldots, a_k \geq 1$ and $a_1+\cdots+a_k=m$. Let M_j denote the number of specimens $m+1, \ldots, m+n$ of species j found in the first sample. Then

$$K_{mn} = \sum_{j=1}^{S_1(m)} \mathbb{1}(M_j > 0) := \sum_{j=1}^{S_1(m)} \xi'_j,$$

where $\xi'_j = \mathbb{1}(M_j > 0)$. From (38) it follows that

$$\mathbb{P}(\xi'_j = 0 \mid A_j(m) = a) = \frac{(\theta+m-a)_{(n)}}{(\theta+m)_{(n)}}$$

Given $S_1(m)=k$, it follows that

$$\begin{aligned}
\mathbb{E}(S_1(m) - K_{mn} \mid S_1(m)=k) &= \mathbb{E}\left(\sum_{j=1}^{k} \frac{(\theta+m-A_j(m))_{(n)}}{(\theta+m)_{(n)}} \mid S_1(m)=k \right) \\
&= \mathbb{E}\left(\sum_{r=1}^{m} C_r(m) \frac{(\theta+m-r)_{(n)}}{(\theta+m)_{(n)}} \mid S_1(m)=k \right) \\
&= \sum_{r=1}^{m} \frac{(\theta+m-r)_{(n)}}{(\theta+m)_{(n)}} \mathbb{E}(C_r(m) \mid S_1(m)=k) \\
&= \sum_{r=1}^{m} \frac{(\theta+m-r)_{(n)}}{(\theta+m)_{(n)}} \frac{m! \left[{m-r \atop k-1} \right]}{r\,(m-r)! \left[{m \atop k} \right]} \\
&= \frac{m!}{(\theta+m)_{(n)} \left[{m \atop k} \right]} \sum_{r=1}^{m} \frac{(\theta+m-r)_{(n)} \left[{m-r \atop k-1} \right]}{r\,(m-r)!},
\end{aligned} \tag{39}$$

the last-but-one line coming from [29, (2.30)], which shows that

$$\mathbb{E}(C_j(m) \mid S(m)=k) = \frac{1}{j} \frac{m!}{\left[{m \atop k} \right]} \frac{\left[{m-j \atop k-1} \right]}{(m-j)!}, j = 1, 2, \ldots, m-k+1.$$

We can use this to compute $\mathbb{E}K_{mn}$ given in (8) in another way, obtaining

Lemma 6.

$$\mathbb{E}K_{mn} = \theta n \sum_{j=0}^{m-1} \frac{1}{(\theta + j)(\theta + n + j)}.$$

Proof. First, we show that

$$\mathbb{E}K_{mn} = \mathbb{E}S_1(m) - \frac{m!\theta}{\theta_{(m+n)}} \sum_{r=1}^{m} \frac{1}{r} \frac{\theta_{(m+n-r)}}{(m-r)!} \tag{40}$$

To establish (40), we average (39) over the distribution of $S_1(m)$ to obtain

$$\begin{aligned}
\mathbb{E}K_{mn} &= \mathbb{E}S_1(m) - \sum_{k=1}^{m} \frac{\theta^k \left[{m \atop k}\right]}{\theta_{(m)}} \frac{m!}{(\theta + m)_{(n)} \left[{m \atop k}\right]} \sum_{r=1}^{m} \frac{(\theta + m - r)_{(n)} \left[{m-r \atop k-1}\right]}{r\,(m-r)!} \\
&= \mathbb{E}S_1(m) - \frac{m!}{\theta_{(m+n)}} \sum_{r=1}^{m} \frac{(\theta + m - r)_{(n)}}{r\,(m-r)!} \sum_{k=1}^{m} \theta^k \left[{m-r \atop k-1}\right] \\
&= \mathbb{E}S_1(m) - \frac{m!}{\theta_{(m+n)}} \sum_{r=1}^{m} \frac{(\theta + m - r)_{(n)} \theta\, \theta_{(m-r)}}{r\,(m-r)!} \\
&= \mathbb{E}S_1(m) - \frac{\theta m!}{\theta_{(m+n)}} \sum_{r=1}^{m} \frac{\theta_{(m+n-r)}}{r\,(m-r)!}.
\end{aligned}$$

This provides a simple formula for the sum on the right of (40). We saw in (8) that

$$\mathbb{E}K_{mn} = \sum_{r=0}^{n-1} \frac{\theta}{\theta + r} - \sum_{r=0}^{n-1} \frac{\theta}{\theta + m + r}$$

Comparing this to (40), we see that

$$\begin{aligned}
\frac{m!\theta}{\theta_{(m+n)}} \sum_{r=1}^{m} \frac{1}{r} \frac{\theta_{(m+n-r)}}{(m-r)!} &= \sum_{r=0}^{m-1} \frac{\theta}{\theta + r} - \sum_{r=0}^{n-1} \frac{\theta}{\theta + r} + \sum_{r=0}^{n-1} \frac{\theta}{\theta + m + r} \\
&= \sum_{r=n}^{n+m-1} \frac{\theta}{\theta + r},
\end{aligned} \tag{41}$$

so that

$$\mathbb{E}K_{mn} = \sum_{r=0}^{m-1} \frac{\theta}{\theta + r} - \sum_{r=n}^{n+m-1} \frac{\theta}{\theta + r},$$

which completes the proof. □

Next we evaluate $\mathbb{E}S_1(m)\,K_{mn}$.

Lemma 7.

$$\mathbb{E}S_1(m)\,K_{mn} = \mathbb{E}(S_1(m))^2 - \sum_{r=n}^{n+m-1} \frac{\theta}{\theta+r} - \frac{\theta m!}{\theta_{(m+n)}} \sum_{r=1}^{m} \frac{1}{r} \frac{\theta_{(m+n-r)}}{(m-r)!} \sum_{i=0}^{m-r-1} \frac{\theta}{\theta+i}. \tag{42}$$

Proof. Multiplying the identity (39) and averaging over the distribution of $S_1(m)$, we have

$$\begin{aligned}\mathbb{E}(S_1(m)K_{mn}) &= \mathbb{E}(S_1(m))^2 - \sum_{k=1}^{m} \frac{\theta^k \begin{bmatrix} m \\ k \end{bmatrix}}{\theta_{(m)}}\, k \sum_{r=1}^{m} \frac{(\theta+m-r)_{(n)}}{(\theta+m)_{(n)}} \frac{m!\begin{bmatrix} m-r \\ k-1 \end{bmatrix}}{r(m-r)!\begin{bmatrix} m \\ k \end{bmatrix}} \\ &= \mathbb{E}(S_1(m))^2 - \frac{m!}{\theta_{(m+n)}} \sum_{r=1}^{m} \frac{(\theta+m-r)_{(n)}}{r(m-r)!} \sum_{k=1}^{m} k\theta^k \begin{bmatrix} m-r \\ k-1 \end{bmatrix}. \end{aligned} \tag{43}$$

The inner sum is

$$\begin{aligned}\theta \sum_{k=1}^{m-r+1} k\theta^{k-1} \begin{bmatrix} m-r \\ k-1 \end{bmatrix} &= \theta \frac{\partial}{\partial\theta} \sum_{k=1}^{m-r+1} \theta^k \begin{bmatrix} m-r \\ k-1 \end{bmatrix} \\ &= \theta \frac{\partial}{\partial\theta} (\theta \cdot \theta_{(m-r)}) \\ &= \theta \left\{ \theta \frac{\partial \theta_{(m-r)}}{\partial\theta} + \theta_{(m-r)} \right\} \\ &= \theta\, \theta_{(m-r)} \left\{ 1 + \sum_{i=0}^{m-r-1} \frac{\theta}{\theta+i} \right\}. \end{aligned} \tag{44}$$

Substituting (44) into (43), we get

$$\mathbb{E}(S_1(m)K_{mn}) = \mathbb{E}(S_1(m))^2 - \frac{\theta m!}{\theta_{(m+n)}} \sum_{r=1}^{m} \frac{\theta_{(m+n-r)}}{r(m-r)!} \left\{ 1 + \sum_{i=0}^{m-r-1} \frac{\theta}{\theta+i} \right\}, \tag{45}$$

which reduces to (42) using the identity (41), as was to be shown. □

A.2. Proof of Lemma 1

Proof. First, we estimate the third term on the right of (9), by proving that

$$\frac{\theta m!}{\theta_{(m+n)}} \sum_{r=1}^{m} \frac{\theta_{(m+n-r)}}{(m-r)!} \frac{1}{r} \sum_{i=m-r}^{m-1} \frac{\theta}{\theta+i} = \frac{\theta^2}{n} + O(m^{-2}). \tag{46}$$

To see this, we rewrite the inner sum as

$$\begin{aligned}\frac{1}{r} \sum_{i=1}^{r} \frac{\theta}{\theta+m-i} &= \frac{1}{r} \left(\frac{r\theta}{\theta+m-1} + \sum_{i=1}^{r} \left(\frac{\theta}{\theta+m-i} - \frac{\theta}{\theta+m-1} \right) \right) \\ &= \frac{\theta}{\theta+m-1} + \frac{\theta}{r(\theta+m-1)} \sum_{i=1}^{r} \frac{i-1}{\theta+m-i}. \end{aligned}$$

Since

$$\frac{r(r-1)}{2(\theta+m-1)} \le \sum_{i=1}^{r} \frac{i-1}{\theta+m-i} \le \frac{r(r-1)}{2(\theta+m-r)},$$

we have

$$\frac{1}{r}\sum_{i=m-r}^{m-1} \frac{\theta}{\theta+i} = \frac{\theta}{\theta+m-1} + \frac{\theta(r-1)}{2(\theta+m-1)^2} + \delta_r, \tag{47}$$

where

$$0 \le \delta_r \le \frac{\theta(r-1)}{2(\theta+m-1)(\theta+m-r)}. \tag{48}$$

Using the fact that for any b and integer k, $\sum_{j=0}^{k} b_{(j)}/j! = (b+1)_{(k)}/k!$, it follows that

$$\frac{\theta m!}{\theta_{(m+n)}} \sum_{r=1}^{m} \frac{\theta_{(m+n-r)}}{(m-r)!} = \frac{\theta m}{\theta+n} \tag{49}$$

It then follows from (49) and (47) that

$$\frac{\theta m!}{\theta_{(m+n)}} \sum_{r=1}^{m} \frac{\theta_{(m+n-r)}}{(m-r)!} \frac{1}{r} \sum_{i=m-r}^{m-1} \frac{\theta}{\theta+i} =$$

$$\frac{m}{\theta+n}\frac{\theta^2}{\theta+m-1} + \frac{\theta m!}{\theta_{(m+n)}} \sum_{r=1}^{m} \frac{\theta_{(m+n-r)}}{(m-r)!} \left(\frac{\theta(r-1)}{2(\theta+m-1)^2} + \delta_r \right).$$

Now, note that there exist $b, b' > 0$ such that for any $m \ge 2$,

$$\frac{\theta^2 b'}{2(\theta+m+n-1)(\theta+m+n-2)} \le \frac{\theta m!}{\theta_{(m+n)}} \sum_{r=1}^{m} \frac{\theta_{(m+n-r)}}{(m-r)!} \frac{\theta(r-1)}{2(\theta+m-1)^2}$$

$$\le \frac{\theta^2 b}{2(\theta+m+n-1)(\theta+m+n-2)} \sum_{r=1}^{\infty} (r-1) \left(\frac{m}{\theta+m+n-1} \right)^{r-2}$$

$$= \frac{\theta^2 b}{2(\theta+m+n-1)(\theta+m+n-2)} \left(\frac{\theta+m+n-1}{\theta+n-1} \right)^2,$$

where the first inequality follows from considering the first two terms of the sum, and the second inequality follows from $(m-r)/(\theta+m+n-r-1) \le m/(\theta+m+n-1)$, for $r \le m$. Finally, using (48) and finding similar bounds for the term involving δ_r, we get

$$\frac{\theta m!}{\theta_{(m+n)}} \sum_{r=1}^{m} \frac{\theta_{(m+n-r)}}{(m-r)!} \frac{1}{r} \sum_{i=m-r}^{m-1} \frac{\theta}{\theta+i} = \frac{m}{\theta+n}\frac{\theta^2}{\theta+m-1} + O(m^{-2}),$$

which completes the proof of (46).

The next step is to estimate the second term on the right-hand side of (9). We show

$$-\sum_{r=n}^{m+n-1}\frac{\theta}{\theta+r}=\theta\log\left(\frac{n}{m+n}\right)+\frac{\theta(\theta-1/2)m}{n(m+n)}+O(m^{-2}). \tag{50}$$

To establish this, using the Euler-Maclaurin summation formula, we can write

$$-\sum_{r=n}^{m+n-1}\frac{\theta}{\theta+r}=\theta\log\left(\frac{\theta+n-1}{\theta+m+n-1}\right)+\frac{\theta m}{2(\theta+n-1)(\theta+m+n-1)}+O(m^{-2}). \tag{51}$$

On the other hand, from the Taylor expansion of $\log(1+x)$

$$\begin{aligned}\log\left(\frac{\theta+n-1}{\theta+m+n-1}\right)-\log\left(\frac{n}{m+n}\right)=&\log\left(1+\frac{\theta-1}{n}\right)-\log\left(1+\frac{\theta-1}{m+n}\right)\\ =&\frac{\theta-1}{n}-\frac{\theta-1}{m+n}+O(m^{-2})\\ =&\frac{(\theta-1)m}{n(m+n)}+O(m^{-2}).\end{aligned} \tag{52}$$

Thus, (51) and (52) imply (50). The lemma now follows by combining (46) and (50). □

A.3. Proof of Lemmas 3 and 4

To establish Lemma 3, we have

$$\begin{aligned}\mathbb{E}[\Psi_n^k]=&\mathbb{E}\left(\int_{\mathbb{R}_+}x^2\nu_n(dx)-\left(\int_{\mathbb{R}_+}x\nu_n(dx)\right)^2\right)^k\\ =&\sum_{i=0}^k(-1)^i\binom{k}{i}\mathbb{E}\left[\left(\int_{\mathbb{R}_+}x^2\nu_n(dx)\right)^{k-i}\left(\int_{\mathbb{R}_+}x\nu_n(dx)\right)^{2i}\right]\\ =&\sum_{i=0}^k(-1)^i\binom{k}{i}\mathbb{E}[\prod_{j=1}^{k-i}S_j^2\prod_{j=1}^{i}S_{k-i+2j-1}S_{k-i+2j}]\\ =&\sum_{i=0}^k(-1)^i\binom{k}{i}\frac{1}{2^k}\mathbb{E}[\prod_{j=1}^{k-i}(S_{2j-1}^2+S_{2j}^2)\prod_{j=k-i+1}^{k}2S_{2j-1}S_{2j}]\\ =&\mathbb{E}[\prod_{i=1}^k\frac{1}{2}(S_{2i-1}-S_{2i})^2].\end{aligned}$$

□

To establish Lemma 4, note that for $p\geq 2k$, it follows from weak exchangeability of $(\phi_{ij})_{i\neq j\in\mathbb{N}}$ that the number of terms in the expansion of $\mathbb{E}[(\sum_{i\neq j}\phi_{ij})^k]$ which are equal to $\mathbb{E}[\prod_{i=1}^k\phi_{2i-1,2i}]$ is

$p(p-1)\cdots(p-2k+1) \sim (p(p-1))^k$, for large p. Recalling that a Skellam distribution always has finite moments of all orders, this implies that

$$\mathbb{E}[\Phi_p^k] \to \mathbb{E}[\prod_{i=1}^{k} \phi_{2i-1,2i}] \leq \max_{i=1,\cdots,k} \mathbb{E}[\phi_{2i-1,2i}^k] = \mathbb{E}[\phi_{1,2}^k] = \frac{1}{2^k}\mathbb{E}[\varphi_{1,2}^{2k}] < \infty,$$

as $p \to \infty$, where the central inequality follows from Hölder's inequality. Therefore, [6, Corollary to Theorem 25.12] gives the moment equality in the statement of Lemma 4. In order to see that the distribution of Φ is determined by its moments, it suffices to show there exists $s > 0$ such that $\mathbb{E}[\exp\{s\Phi\}] < \infty$ [6, Theorem 30.1]. Note that, from exchangeability and Hölder's inequality, there exists $s > 0$ such that for any $p \geq 2$

$$\mathbb{E}[e^{s\Phi_p}] \leq \max_{i \neq j \leq p} \mathbb{E}[e^{s\phi_{ij}}] = \mathbb{E}[e^{s\phi_{1,2}}] = \mathbb{E}[e^{\frac{s}{2}\varphi_{1,2}^2}] < \infty.$$

Thus, from the continuous mapping theorem and Fatou's lemma, we deduce

$$\mathbb{E}[e^{s\Phi}] \leq \liminf_p \mathbb{E}[e^{s\phi_p}] \leq \mathbb{E}[e^{\frac{s}{2}\varphi_{1,2}^2}] < \infty.$$

□

A.4. Proofs of Propositions 2 and 3

Proof of Proposition 2. The expected value of Φ may be easily obtained from Lemma 4. Applying Lemma 4 for the variance, note that $\lim_{p\to\infty} \operatorname{var}\Phi_p = \operatorname{var}(\Phi) = \mathbb{E}\phi_{12}\phi_{34} - (\mathbb{E}\phi_{12})^2 = \operatorname{cov}(\phi_{12}, \phi_{34})$. Hence, it suffices to prove

$$\operatorname{cov}(\phi_{i,j}, \phi_{k,l}) = \operatorname{cov}(\phi_{1,2}, \phi_{3,4}) = \theta \log(9/8).$$

From the symmetry of π_{ij}

$$\operatorname{cov}(\phi_{1,2}, \phi_{3,4}) = \operatorname{cov}(\pi_{12}^2, \pi_{34}^2) - 2\operatorname{cov}(\pi_{12}^2, \pi_{34}\pi_{43}) + \operatorname{cov}(\pi_{12}\pi_{21}, \pi_{34}\pi_{43}). \tag{53}$$

In order to compute the right hand side of (53), for $1 \leq i \neq j \leq 4$, we can write from (22) and (23)

$$\pi_{ij} = \overline{\pi}_i^{(4)} + \overline{\pi}_{ir}^{(4)} + \overline{\pi}_{is}^{(4)} + \overline{\pi}_{irs}^{(4)},$$

where $\{r \neq s\} = \{1,2,3,4\} \setminus \{i,j\}$. To ease the notation, in the sequel, we drop the superscript (4) from $\overline{\pi}_A$. From the symmetry and independence of $\overline{\pi}_A$, we obtain

$$\begin{aligned}
\operatorname{cov}(\pi_{12}^2, \pi_{34}^2) &= \operatorname{cov}(\overline{\pi}_{13}^2 + 2\overline{\pi}_{13}(\overline{\pi}_1 + \overline{\pi}_{14} + \overline{\pi}_{134}), \overline{\pi}_{13}^2 + 2\overline{\pi}_{13}(\overline{\pi}_3 + \overline{\pi}_{23} + \overline{\pi}_{123})) \\
&= \operatorname{var}\overline{\pi}_{13}^2 + 4\operatorname{cov}(\overline{\pi}_{13}^2, \overline{\pi}_{13}(\overline{\pi}_3 + \overline{\pi}_{23} + \overline{\pi}_{123})) \\
&\quad + 4\operatorname{cov}(\overline{\pi}_{13}(\overline{\pi}_1 + \overline{\pi}_{14} + \overline{\pi}_{134}), \overline{\pi}_{13}(\overline{\pi}_3 + \overline{\pi}_{23} + \overline{\pi}_{123})) \\
&= \mathbb{E}\overline{\pi}_{13}^4 - (\mathbb{E}\overline{\pi}_{13}^2)^2 + 4(\mathbb{E}\overline{\pi}_{13}^3 - \mathbb{E}\overline{\pi}_{13}^2\mathbb{E}\overline{\pi}_{13})\mathbb{E}[\overline{\pi}_3 + \overline{\pi}_{23} + \overline{\pi}_{123}] \\
&\quad + 4\operatorname{var}(\overline{\pi}_{13})(\mathbb{E}[\overline{\pi}_1 + \overline{\pi}_{14} + \overline{\pi}_{134}])^2.
\end{aligned} \tag{54}$$

Similarly, one can derive

$$\begin{aligned}\operatorname{cov}(\pi_{12}^2, \pi_{34}\pi_{43}) = {} & 2(\mathbb{E}\overline{\pi}_{13}^3 - \mathbb{E}\overline{\pi}_{13}^2\mathbb{E}\overline{\pi}_{13})(\mathbb{E}\overline{\pi}_{14} + \mathbb{E}[\overline{\pi}_4 + \overline{\pi}_{24} + \overline{\pi}_{124}]) \\ & + 4\operatorname{var}(\overline{\pi}_{13})\mathbb{E}[\overline{\pi}_4 + \overline{\pi}_{24} + \overline{\pi}_{124}](\mathbb{E}\overline{\pi}_{14} + \mathbb{E}[\overline{\pi}_1 + \overline{\pi}_{134}]) \\ & + 4\operatorname{var}(\overline{\pi}_{13})\mathbb{E}[\overline{\pi}_1 + \overline{\pi}_{134}]\mathbb{E}\overline{\pi}_{14} + 2(\mathbb{E}\overline{\pi}_{13}^2)^2 - 2(\mathbb{E}\overline{\pi}_{13})^4 \end{aligned} \tag{55}$$

and

$$\begin{aligned}\operatorname{cov}(\pi_{12}\pi_{21}, \pi_{34}\pi_{43}) = {} & 12\operatorname{var}(\overline{\pi}_{13})(\mathbb{E}\overline{\pi}_{13})^2 + 16\operatorname{var}(\overline{\pi}_{13})\mathbb{E}[\overline{\pi}_2 + \overline{\pi}_{234}]\mathbb{E}\overline{\pi}_{14} \\ & + 2\operatorname{var}(\overline{\pi}_{13}\overline{\pi}_{24}) + 4\operatorname{var}(\overline{\pi}_{13})(\mathbb{E}[\overline{\pi}_2 + \overline{\pi}_{234}])^2. \end{aligned} \tag{56}$$

Substituting (54),(55) and (56) into (53) and simplifying, we get

$$\operatorname{cov}(\phi_{1,2}, \phi_{3,4}) = \mathbb{E}\overline{\pi}_{13} = \tilde{\lambda}_2^{(4)} = \theta \log(9/8). \qquad \square$$

Proof of Proposition 3. We write α_p in the form

$$\begin{aligned}\alpha_p &= \sum_{r=1}^{p-1}(-1)^r \binom{p-1}{r} \log r - \sum_{r=1}^{p-1}(-1)^r \binom{p-1}{r} \log(1+r) \\ &=: \alpha_{1p} - \alpha_{2p}\end{aligned}$$

First, consider α_{2p}. Note that

$$\begin{aligned}\binom{p-1}{r} &= \frac{(p-1)!}{(p-1-r)!r!} = \frac{1+r}{p}\,\frac{p!}{(p-[r+1])!(r+1)!} \\ &= \frac{1+r}{p}\binom{p}{r+1},\end{aligned}$$

so that

$$\begin{aligned}\alpha_{2p} &= \frac{1}{p}\sum_{r+1=2}^{p}(-1)^{r+1-1}\binom{p}{r+1}(r+1)\log(r+1) \\ &= \frac{1}{p}\sum_{l=2}^{p}(-1)^{l-1}\binom{p}{l} l\log(l) \\ &= -\frac{1}{p}\sum_{l=1}^{p}(-1)^{l}\binom{p}{l} l\log(l)\end{aligned}$$

We can use [7], which shows that as $p \to \infty$

$$\sum_{k=1}^{p}(-1)^k\binom{p}{k} k\log k = \frac{1}{\log p} + O(1/(\log p)^2)$$

to see that

$$\alpha_{2p} = -\frac{1}{p\log p} + O(1/(p\log^2 p)).$$

For α_{1p} we have the result from [15] that gives

$$\alpha_{1p} = \log\log(p-1) + \gamma + \frac{\gamma}{\log(p-1)} + O(1/\log^2(p-1)).$$

Putting it together, we get

$$\alpha_p = \log\log p + \gamma + \frac{\gamma}{\log p} + \frac{1}{p\log p} + O(\log^{-2} p),$$

as claimed. □

Acknowledgements

We thank Stephen Senn for bringing the Anscombe-Fisher correspondence to our attention. We thank two reviewers and the associate editor for comments that improved the paper. PHdS and ST were supported in part by NSF grant DMS-2030562.

References

[1] Aldous, D.J. (1985). Exchangeability and related topics. In *École D'été de Probabilités de Saint-Flour, XIII—1983*. *Lecture Notes in Math.* **1117** 1–198. Berlin: Springer. MR0883646 https://doi.org/10.1007/BFb0099421

[2] Anscombe, F.J. (1950). Sampling theory of the negative binomial and logarithmic series distributions. *Biometrika* **37** 358–382. MR0039193 https://doi.org/10.1093/biomet/37.3-4.358

[3] Arratia, R., Barbour, A.D. and Tavaré, S. (2003). *Logarithmic Combinatorial Structures: A Probabilistic Approach*. *EMS Monographs in Mathematics*. Zürich: European Mathematical Society (EMS). MR2032426 https://doi.org/10.4171/000

[4] Arratia, R. and Tavaré, S. (1992). Limit theorems for combinatorial structures via discrete process approximations. *Random Structures Algorithms* **3** 321–345. MR1164844 https://doi.org/10.1002/rsa.3240030310

[5] Barbour, A.D. and Tavaré, S. (2010). Assessing molecular variability in cancer genomes. In *Probability and Mathematical Genetics*. *London Mathematical Society Lecture Note Series* **378** 91–112. Cambridge: Cambridge Univ. Press. MR2744236

[6] Billingsley, P. (1995). *Probability and Measure*, 3rd ed. *Wiley Series in Probability and Mathematical Statistics*. New York: Wiley. MR1324786

[7] Chapling, R. (2016). Asymptotics of certain sums required in loop regularisation. *Modern Phys. Lett. A* **31** 1650030. MR3454725 https://doi.org/10.1142/S0217732316500309

[8] Crane, H. (2016). The ubiquitous Ewens sampling formula. *Statist. Sci.* **31** 1–19. MR3458585 https://doi.org/10.1214/15-STS529

[9] Eagleson, G.K. and Weber, N.C. (1978). Limit theorems for weakly exchangeable arrays. *Math. Proc. Cambridge Philos. Soc.* **84** 123–130. MR0501275 https://doi.org/10.1017/S0305004100054967

[10] Engen, S. (1978). *Stochastic Abundance Models*. *Monographs on Applied Probability and Statistics*. London: Chapman and Hall. MR0515721

[11] Ewens, W.J. (1972). The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.* **3** 87–112. MR0325177 https://doi.org/10.1016/0040-5809(72)90035-4

[12] Ewens, W.J., Choudhury, A.R., Lewontin, R.C. and Wiuf, C. (2007). Two variance results in population genetics theory. *Math. Popul. Stud.* **14** 93–110. MR2316681 https://doi.org/10.1080/08898480701298376

[13] Fisher, R.A. (1943). A theoretical distribution for the apparent abundance of different species. *J. Anim. Ecol.* **12** 54–57.

[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. Anim. Ecol.* **12** 42–58.

[15] Flajolet, P. and Sedgewick, R. (1995). Mellin transforms and asymptotics: Finite differences and Rice's integrals. *Theoret. Comput. Sci.* **144** 101–124. MR1337755 https://doi.org/10.1016/0304-3975(94)00281-M
[16] Irwin, J.O. (1937). The frequency distribution of the difference between two independent variates following the same Poisson distribution. *J. Roy. Statist. Soc. Ser. A* **100** 415–416.
[17] Johnson, N.L., Kotz, S. and Balakrishnan, N. (1997). *Discrete Multivariate Distributions*. *Wiley Series in Probability and Statistics: Applied Probability and Statistics*. New York: Wiley. MR1429617
[18] Kallenberg, O. (2006). *Foundations of Modern Probability*. Springer Science & Business Media.
[19] Kendall, D.G. (1948). On some modes of population growth leading to R. A. Fisher's logarithmic series distribution. *Biometrika* **35** 6–15. MR0026282 https://doi.org/10.1093/biomet/35.1-2.6
[20] Lijoi, A., Mena, R.H. and Prünster, I. (2007). Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika* **94** 769–786. MR2416792 https://doi.org/10.1093/biomet/asm061
[21] Lijoi, A., Prünster, I. and Walker, S.G. (2008). Bayesian nonparametric estimators derived from conditional Gibbs structures. *Ann. Appl. Probab.* **18** 1519–1547. MR2434179 https://doi.org/10.1214/07-AAP495
[22] Lindsay, B.G. and Basak, P. (2000). Moments determine the tail of a distribution (but not much else). *Amer. Statist.* **54** 248–251. MR1803622 https://doi.org/10.2307/2685775
[23] Lindsay, B.G., Pilla, R.S. and Basak, P. (2000). Moment-based approximations of distributions using mixtures: Theory and applications. *Ann. Inst. Statist. Math.* **52** 215–230. MR1763559 https://doi.org/10.1023/A:1004105603806
[24] McCullagh, P. (2016). Two early contributions to the Ewens saga [comment on MR3458585]. *Statist. Sci.* **31** 23–26. MR3458587 https://doi.org/10.1214/15-STS536
[25] Pitman, J. (2006). *Combinatorial Stochastic Processes*. *Lecture Notes in Math.* **1875**. Berlin: Springer. MR2245368
[26] Sibuya, M. (2014). Prediction in Ewens-Pitman sampling formula and random samples from number partitions. *Ann. Inst. Statist. Math.* **66** 833–864. MR3250819 https://doi.org/10.1007/s10463-013-0427-8
[27] Skellam, J.G. (1946). The frequency distribution of the difference between two Poisson variates belonging to different populations. *J. Roy. Statist. Soc. (N.S.)* **109** 296. MR0020750
[28] Tavaré, S. (2021). The magical Ewens sampling formula. *Bull. Lond. Math. Soc.* **53** 1563–1582. MR4368686 https://doi.org/10.1112/blms.12537
[29] Watterson, G.A. (1974). Models for the logarithmic species abundance distributions. *Theor. Popul. Biol.* **6** 217–250. MR0368829 https://doi.org/10.1016/0040-5809(74)90025-2

Received October 2021 and revised March 2022