KARLIN-MCGREGOR MUTATIONAL OCCUPANCY PROBLEM REVISITED

Thierry Huillet

KARLIN-MCGREGOR MUTATIONAL OCCUPANCY PROBLEM REVISITED

THIERRY E. HUILLET

Abstract. Some population is made of n individuals that can be of p possible species (or types). The update of the species abundance occupancies is from a Moran mutational model designed by Karlin and McGregor in 1967. We first study the equilibrium species counts as a function of n, p and the total mutation probability ν before considering various asymptotic regimes on n, p and ν.

Running title: KMG Model with Mutations.

Keywords: Species abundance; Karlin-McGregor-Moran Models; Mutational and evolutionary processes; Population dynamics. Asymptotics.

1. Introduction

Some population is made of n individuals that can be of p possible species (or types). The discrete-time update of the species abundance occupancies is from a Moran mutational model first designed in [9] and for which the size n of the population is maintained constant over the generations. We will study in great detail the equilibrium species counts as a function of n, p and the total mutation probability ν before considering various asymptotic regimes of interest on n, p and ν, some of which were not considered in [9]. When they exist while $n \wedge p \to \infty$, the limiting distributions of the typical species abundance are not heavy-tailed, rather they have a dominant exponential decay factor and this may be seen to result from the conservation of the population size n. They are rather related to the negative binomial or Fisher log-series distributions, [5]. Also of particular interest will be (i) the distribution of the number of occupied species with positive occupancy (ii) the probability that two randomly sampled individuals are of the same species; this both for fixed n, p and ν and under their asymptotic regimes.

This model should not be confused with the following related (although non-conservative) Yule mutation model, [19], [21]: A species starts with a single individual. As a result of mutations, new individuals are produced according to a linear pure birth Yule process with some birth rate and they all belong to the same species. Concomitantly and as a result of specific mutations, inside a species, an individual of a novel species can be created at some other rate and the new species, once it has appeared, behaves like all the previous ones. For the Yule model, the asymptotic abundance inside a typical species is distributed like a Simon distribution [17] which (in sharp contrast with the former log-series-like distribution), is heavy-tailed, translating the presence of very large family counts. Note that here

both n and the number p of possible species should be set to infinity because both are bound to grow indefinitely in the process, see [16].

The Karlin-McGregor (KMG) mutation model was originally developed to study multiallelic frequencies dynamics in population genetics, as from [11]. It was later applied to the study of surname distributions and random isonymy, making the observation that surnames can be considered as alleles transmitted along the male line. See [20], [12], [22], [15] and the references therein. One can apply the model not only to surnames (which can be linked to $Y-$chromosomes) but also to first names and other elements of culture that do propagate by copying.

2. Species abundances evolution: the KMG mutation model

Some population is made of n individuals that can be of p possible species (or types).

At (discrete-time) step t, there are $K_t(q) \geq 0$ individuals of type q, $q = 1, ..., p$. The occupancy vector $\mathbf{K}_t := (K_t(q); q = 1, ..., p)$ is called the species abundance vector. The species q will be said filled if $K_t(q) > 0$ (it has at least one representative).

We let $Q_t := \sum_{q=1}^{p} \mathbf{1}(K_t(q) > 0)$ be the number of types present at step t (the number of filled species).

We let $N_t(k) := \sum_{q=1}^{p} \mathbf{1}(K_t(q) = k)$ be the number of species with k representatives at step t.

We have $1 \leq Q_t = p - N_t(0) \leq p \wedge n$ and $\sum_{q=1}^{p} K_t(q) = n = \sum_{k=1}^{\max_q K_t(q)} kN_t(k)$.

The dynamics of $\mathbf{K}_t$ is in the spirit of a Moran $\beta-$mutation evolution model, preserving the total number of individuals n, namely, [13], [7]:

Given $K_t(q) = k_q$, $q = 1, ..., p$, we let $(k_1, ..., k_p) \to (k_1, ..., k_q - 1, ..., k_{q'} + 1, ..., k_p)$ be the moves between step t and step $t+1$: at each step, an individual of type q is deleted from the population and an individual of type $q' \neq q$ is created. We assume that this event occurs with probability (w.p.)

$$\frac{k_q}{n}\left[\frac{k_{q'}}{n}(1-(p-1)\beta) + \left(1-\frac{k_{q'}}{n}\right)\beta\right]. \tag{1}$$

For such a mutation model, an individual of type q is deleted (with probability $\frac{k_q}{n}$) and an individual of type q' is created either because q' is selected to duplicate (with probability $\frac{k_{q'}}{n}$) and the duplicate has not mutated to any other state than q' (an event of probability $1-(p-1)\beta$) or because an individual of type $q'' \neq q'$ is selected to duplicate (with probability $1-\frac{k_{q'}}{n}$) and the duplicate has mutated to an individual of type q' (with probability β). We let $p\beta = \nu$ be the overall mutation probability.

When the $K_{t=0}(q)$'s are exchangeable, the $K_t(q)$'s remain exchangeable for all t (having law invariant upon a permutation of the q's), in particular all the $K_t(q)$'s share the same distribution. Let us thus focus on $K_t(1)$ with $K_t(q) \overset{d}{=} K_t(1)$, $q = 2, ..., p$ (equality in distribution). Then, see [9], while lumping the states $K_t(q)$, $q = 2, ..., n$, given $K_t(1) = k \in \{0, ..., n\}$

$$\begin{array}{l}(k, n-k) \to (k+1, n-k-1) \text{ w.p. } p_k = \left(1-\frac{k}{n}\right)\left(\frac{k}{n}(1-(p-1)\beta) + \left(1-\frac{k}{n}\right)\beta\right) \\ (k, n-k) \to (k-1, n-k+1) \text{ w.p. } q_k = \frac{k}{n}\left(\frac{k}{n}(p-1)\beta + \left(1-\frac{k}{n}\right)(1-\beta)\right)\end{array}$$

defines the tridiagonal transition probabilities of a random walk on the set $\{0,...,n\}$ with holding probability $r_k = 1-(p_k+q_k)$ that $(k,n-k) \to (k,n-k)$. This random walk is ergodic with invariant probability measure (independent of the initial condition $K_{t=0}(1)$) given for $k=0,...,n$ by (see [10] or [7] for instance):

$$\pi_k := \mathbf{P}(K_\infty(1)=k) = \frac{\binom{k+\frac{1}{p}n\nu/(1-\nu)-1}{k}\binom{n/(1-\nu)-k-\frac{1}{p}n\nu/(1-\nu)-1}{n-k}}{\binom{n/(1-\nu)-1}{n}}. \tag{2}$$

This is also

$$\begin{array}{l}\pi_k = \binom{n}{k}\frac{B(k+\theta,n-k+(p-1)\theta)}{B(\theta,(p-1)\theta)} \\ = \binom{n}{k}\frac{\Gamma(n\nu/(1-\nu))}{\Gamma(n/(1-\nu))}\frac{\Gamma(k+\theta)}{\Gamma(\theta)}\frac{\Gamma(n/(1-\nu)-k-\theta)}{\Gamma(n\nu/(1-\nu)-\theta)}\end{array} \tag{3}$$

where $\theta = \frac{n}{p}\nu/(1-\nu)$ and $B(a,b)$ is the beta function. In particular, $\pi_0 = \frac{\Gamma(n\nu/(1-\nu))}{\Gamma(n/(1-\nu))}\frac{\Gamma(n/(1-\nu)-\theta)}{\Gamma(n\nu/(1-\nu)-\theta)}$. The distribution π_k of $K(1) := K_\infty(1)$ is a $B(\theta,(p-1)\theta)$ s-mixture of a binomial bin(n,s) distribution, $s\in(0,1)$. It is a Pólya-Eggenberger distribution with probability generating function (pgf)

$$\mathbf{E}\left(u^{K(1)}\right) = F(-n,\theta;p\theta;1-u),$$

where $F := {}_2F_1$ is a Gauss hypergeometric function. One can check that $K(1)$ has mean $\mathbf{E}(K(1)) = n/p$ and variance

$$\sigma^2(K(1)) = \frac{n(p-1)(n+p\theta)}{p^2(p\theta+1)} = \left(\frac{n}{p}\right)^2\frac{p-1}{1+\nu(n-1)}.$$

An interesting immediate consequence is the following: noting that $p_k := k\pi_k/\mathbf{E}(K(1))$ is the size-biased probability to pick an individual with k representatives at equilibrium, the probability α that two randomly chosen individuals from the population are of the same species is

$$\alpha = \sum_{k=1}^n \frac{k}{n}p_k = \frac{p}{n^2}\sum_{k=1}^n k^2\pi_k = \frac{p}{n^2}\left(\sigma^2(K(1)) + \mathbf{E}(K(1))^2\right) = \frac{p+\nu(n-1)}{p(1+\nu(n-1))}. \tag{4}$$

The one-dimensional law of $K(1)$ being under control for all n,p, we now wish to evaluate its asymptotic shape under various limiting conditions on n,p, namely $n\approx p$, $n\ll p$ and $n\gg p$ corresponding respectively to $\mu := n/p = O(1)$, $\mu\to 0$ and $\mu\to\infty$. For each asymptotic regime, we shall denote by "*" the asymptotic evaluation of the quantities of interest.

3. Various asymptotics

We shall study five asymptotic regimes depending on the density μ of individuals over the species range.

1. (balanced case)**.** If both $p,n\to\infty$ while $\mu = n/p \to \mu^* > 0$ and ν fixed, then $\theta = \frac{n}{p}\nu/(1-\nu) \sim \theta^* = \mu^*\nu/(1-\nu) > 0$ and

$$\begin{array}{l}\pi_k = \binom{n}{k}\frac{\Gamma(n\nu/(1-\nu))}{\Gamma(n/(1-\nu))}\frac{\Gamma(k+\theta)}{\Gamma(\theta)}\frac{\Gamma(n/(1-\nu)-k-\theta)}{\Gamma(n\nu/(1-\nu)-\theta)} \\ \sim \frac{n^k}{k!}\frac{\Gamma(k+\theta^*)}{\Gamma(\theta^*)}\frac{(n/(1-\nu))^{-(k+\theta^*)}}{(n\nu/(1-\nu))^{-\theta^*}} = \frac{\nu^{\theta^*}}{k!}\frac{\Gamma(k+\theta^*)}{\Gamma(\theta^*)}(1-\nu)^k =: \pi_k^*\end{array} \tag{5}$$

is a well-defined negative binomial distribution for all $k \in \{0,1,2,...\}$, so with limiting pgf $\mathbf{E}\left(u^{K(1)}\right) \sim \left(\nu/\left(1-(1-\nu)\,u\right)\right)^{\theta^*}$. When k is large $\pi_k^* \sim \frac{\nu^{\theta^*}}{k!}\frac{\Gamma(k+\theta^*)}{\Gamma(\theta^*)}\left(1-\nu\right)^k \sim \frac{\nu^{\theta^*}}{\Gamma(\theta^*)}k^{\theta^*-1}\left(1-\nu\right)^k$, a distribution displaying an algebraic prefactor (if $\theta^* \neq 1$) combined to a dominant geometric cutoff. The mean of $K(1)$ is μ^* while its variance is $\mu^*/\nu > \mu^*$ (overdispersion holds). We have

$$\frac{\pi_{k+1}^*}{\pi_k^*} = \frac{k+\theta^*}{k+1}\left(1-\nu\right),$$

so that if $\frac{\pi_1^*}{\pi_0^*} > 1$ $(\mu^*\nu > 1)$, the mode of this distribution is away from zero at about $\left(\mu^*\nu - 1\right)/\nu$; otherwise the mode is at the origin.

The size-biased version of π_k^* is $p_k^* = k\pi_k^*/\mu^*$ and the limiting probability α^* that two randomly chosen individuals from the population are of the same species tends to 0 like

$$\alpha^* := \sum_{k=1}^{n}\frac{k}{n}p_k^* = \frac{1}{n\mu^*}\sum_{k=1}^{n}k^2\pi_k^* = \frac{1}{p}\left(1+\frac{1}{\mu^*\nu}\right).$$

Note that under this asymptotic regime, with $Q = Q_\infty$, the limiting number of species present in the population,

$$\mathbf{E}\left(Q\right) = \sum_{q=1}^{p}\mathbf{P}\left(K\left(q\right)>0\right) = p\left(1-\mathbf{P}\left(K\left(1\right)=0\right)\right) \sim p\left(1-\pi_0^*\right) = n\frac{1-\nu^{\theta^*}}{\mu^*} \to \infty.$$

It scales like a fraction of n because $1-\nu^{\theta^*} < \mu$ as a result of $-\theta^*\log\nu = \mu^*\nu\log\left(1/\nu\right)/\left(1-\nu\right) < -\log\left(1-\mu^*\right)$ and $\log\left(1/\nu\right) < \left(1-\nu\right)/\nu$ for all $\nu \in (0,1)$. Note that as a result, $\pi_0^* > 1-\mu^*$ which is useful only if $\mu^* \in (0,1)$. We will show below that $\sigma^2\left(Q\right) \sim p\left(\nu^{\theta^*} - \nu^{2\theta^*}\right)$. This asymptotic regime was not considered in [9].

2. First fix n. If now as in [9], we let $p \to \infty$ (infinitely many possible types in the population, see [11] for a justification of this in population genetics) and $\beta \to 0$ (small mutation probability) while $p\beta = \nu > 0$ is fixed, then $\mu = \frac{n}{p} \to 0$ and $\theta = \frac{n}{p}\nu/\left(1-\nu\right) \to 0$ while $p\theta \to n\nu/\left(1-\nu\right)$. To the leading order, as $p \to \infty$

$$\begin{array}{l}\pi_k \sim \theta\binom{n}{k}\frac{\Gamma(k)\Gamma(n/(1-\nu)-k)}{\Gamma(n/(1-\nu))} = \pi_k^*,\ k=1,...,n \\ \pi_0 \sim 1-\theta\left(\frac{\Gamma'(n/(1-\nu))}{\Gamma(n/(1-\nu))} - \frac{\Gamma'(n\nu/(1-\nu))}{\Gamma(n\nu/(1-\nu))}\right) = \pi_0^*,\end{array} \tag{6}$$

showing that the equilibrium mass concentrates on state zero: in this low density regime, the number n of individuals being very few compared to p, the typical species occupancy is very low.

However (with $\psi\left(z\right) := \Gamma'\left(z\right)/\Gamma\left(z\right)$ the digamma function), given $K\left(1\right) \geq 1$, for all $k = 1,...,n$,

$$\mathbf{P}\left(K\left(1\right)=k \mid K\left(1\right)\geq 1\right) = \frac{\pi_k^*}{1-\pi_0^*} \sim \binom{n}{k}\frac{B\left(k, n/\left(1-\nu\right)-k\right)}{\psi\left(n/\left(1-\nu\right)\right)-\psi\left(n\nu/\left(1-\nu\right)\right)} \tag{7}$$

is a well-defined probability mass function as $\theta \to 0$ $(p \to \infty)$ and fixed n and ν: given a species is filled, it has a well-defined occupancy distribution. Note in passing that this leads to the non-trivial identity involving the digamma function:

for all $\nu \in (0,1)$

$$\psi\left(\frac{n}{1-\nu}\right)-\psi\left(\frac{n\nu}{1-\nu}\right)=\sum_{k=0}^{n-1}\frac{1}{k+n\nu/(1-\nu)}=\sum_{k=1}^{n}\binom{n}{k}B\left(k,n/(1-\nu)-k\right).$$

This results from (7) and from $\psi(z+1)-\psi(z)=1/z$ so that by telescopic summation: $\psi(z+n)-\psi(z)=\sum_{k=0}^{n-1}1/(k+z)$.

With $(n)_k = n!/(n-k)!$, the mean of π_k^* is

$$\mu=\theta\sum_{k=1}^{n}k\binom{n}{k}\frac{\Gamma(k)\Gamma(n/(1-\nu)-k)}{\Gamma(n/(1-\nu))}=\theta\frac{1-\nu}{\nu}.$$

It vanishes like θ. However, the size-biased version of π_k^*, namely $p_k^*=k\pi_k^*/\mu$, is well-defined, and the limiting probability α^* that two randomly chosen individuals from the population are of the same species is (see 4.20 of [9])

$$\alpha^*:=\sum_{k=1}^{n}\frac{k}{n}p_k^*=\frac{1}{n\mu}\sum_{k=1}^{n}k^2\pi_k^*=\frac{1}{1+\nu(n-1)},$$

which could have been guessed from (4) as $p\to\infty$, ν fixed.

If now n itself tends to ∞ while ν is still held fixed and $n/p\to 0$ (so that θ still tends to 0 as well), owing to $\psi(z)=\frac{\Gamma'(z)}{\Gamma(z)}\underset{z\to\infty}{\sim}\log z$ and the Stirling formula,

$$\mathbf{P}\left(K(1)=k\mid K(1)\geq 1\right)\sim\frac{1}{k}\frac{(1-\nu)^k}{\log(1/\nu)},\ k\geq 1, \tag{8}$$

a Fisher log-series distribution displaying an hyperbolic prefactor combined to a geometric cutoff, [16], [5], [2] and [3]. Note again that $\mu=n/p\to 0$ stipulates that on average each of the species abundances vanish and only given a species is filled, does it has a well-defined occupancy distribution.

In this asymptotic regime, with $Q=Q_\infty$, the limiting number of species present in the population,

$$\mathbf{E}(Q)\sim p(1-\pi_0^*)\sim p\theta\log(1/\nu)=n\nu\log(1/\nu)/(1-\nu)\to\infty \tag{9}$$

and it scales like a fraction of n as well (recalling $\log(1/\nu)<(1-\nu)/\nu$ for all $\nu\in(0,1)$). From [9]

$$\sigma^2(Q)\sim n\left[\nu\log(1/\nu)/(1-\nu)-\nu\right]>0$$

suggesting that $(Q-\mathbf{E}(Q))/\sigma(Q)$ is asymptotically normal. Note $\sigma^2(Q)<\mathbf{E}(Q)$ (underdispersion).

3. Suppose now that $n\to\infty$, $\nu\to 0$ while $n\nu=\lambda>0$ is held fixed and $p\nu\to\infty$. Then $\theta=\frac{n}{p}\nu/(1-\nu)\sim\frac{\lambda}{p}\to 0$ and, with $k=[nx]$

$$\pi_k\sim\frac{\lambda}{p}\binom{n}{nx}\frac{\Gamma(nx)\Gamma(n(1-x)+\lambda)}{\Gamma(n+\lambda)}=\frac{\lambda}{pnx}\frac{\Gamma(n(1-x)+1+\lambda-1)}{\Gamma(n(1-x)+1)}\frac{\Gamma(n+1)}{\Gamma(n+1+\lambda-1)}\sim\frac{\lambda}{pk}\left(1-\frac{k}{n}\right)^{\lambda-1},\ k\geq 1 \tag{10}$$

$$\pi_0\sim 1-\frac{\lambda}{p}\left(\frac{\Gamma'(n)}{\Gamma(n)}-\frac{\Gamma'(\lambda)}{\Gamma(\lambda)}\right)\sim 1-\frac{\lambda}{p}\log n=:\pi_0^*,$$

showing that the equilibrium probability mass concentrates on state zero. Note that since here $\theta\to 0$ ($p\to\infty$) and $n\to\infty$, $\nu\to 0$ while $n\nu=\lambda>0$, then

$\mu = n/p \sim \frac{\theta}{\nu} \sim \frac{\lambda}{p\nu} \to 0$ if $p\nu \to \infty$ (on average each species abundance vanishes). With $x \in (0,1)$ and $k = [nx] \to \infty$, putting $n^{-1} = dx$, we have

$$\pi^*_{[nx]} \sim \frac{\lambda}{p} x^{-1} (1-x)^{\lambda-1} dx,$$

not a probability density. Following [9] however, $p\pi^*_k = \mathbf{E}(N(k)) \sim \frac{\lambda}{k}\left(1-\frac{k}{n}\right)^{\lambda-1}$ is also the asymptotic expected number of mutants in the population with k representatives. This shows that in this regime, the expected number of species whose frequencies range in the interval $(x_1, x_2) \subseteq [0,1]$ is $\lambda \int_{x_1}^{x_2} x^{-1}(1-x)^{\lambda-1} dx$ as $n \to \infty$. Note $\lambda \int_{1/n}^1 x^{-1}(1-x)^{\lambda-1} dx \sim \lambda \log n \sim p(1-\pi_0)$ while $\lambda \int_0^1 x^{-1}(1-x)^{\lambda-1} dx = \infty$.

With $Q = Q_\infty$, the limiting number of species present in the population, this is consistent with

$$\mathbf{E}(Q) = p(1-\pi_0) \sim \lambda \log n - \lambda \frac{\Gamma'(\lambda)}{\Gamma(\lambda)}. \tag{11}$$

In this regime, $\mathbf{E}(Q)$ scales like $\log n$ and (mutations being rare) the asymptotic number of types present is sparse compared to n. It is also shown in [9] that $\sigma^2(Q) \sim \lambda \log n$, so that, upon scaling, $(Q - \mathbf{E}(Q))/\sigma(Q)$ is asymptotically normal.

4. The authors of [9] also consider the asymptotic regime for which $n \to \infty$, $\nu \to 0$ while $\nu n \log n = c \geq 0$ ($\lambda = c/\log n \to 0$) for which from the above estimates and $\lambda \frac{\Gamma'(\lambda)}{\Gamma(\lambda)} \underset{\lambda \to 0^+}{\sim} -1$, $\mathbf{E}(Q) \sim 1 + c$ and $\sigma^2(Q) \sim c$. In this asymptotic regime, only a finite number of types are present.

5. (the dense case). If p is fixed and $n \to \infty$, $\nu \to 0$ while $n\nu = \lambda > 0$, then $\theta = \frac{n}{p}\nu/(1-\nu) \to \theta^* = \lambda/p > 0$ and

$$\begin{array}{l} \pi_k \sim \frac{\Gamma(\lambda)}{\Gamma(\theta^*)\Gamma(\lambda-\theta^*)} \frac{\Gamma(k+1+\theta^*-1)}{\Gamma(k+1)} \frac{\Gamma(n-k+1+\lambda-\theta^*-1)}{\Gamma(n-k+1)} \frac{\Gamma(n+1)}{\Gamma(n+1+\lambda-1)} \\ \sim \frac{\Gamma(\lambda)}{\Gamma(\theta^*)\Gamma(\lambda-\theta^*)} \frac{\Gamma(k+1+\theta^*-1)}{\Gamma(k+1)} \frac{(n-k+1)^{\lambda-\theta^*-1}}{(n+1)^{\lambda-1}} = \pi^*_k. \end{array} \tag{12}$$

If $k = [nx] \to \infty$ with $x \in (0,1)$

$$\begin{array}{l} \pi^*_{[nx]} \sim n^{-1} n \frac{\Gamma(\lambda)}{\Gamma(\theta^*)\Gamma(\lambda-\theta^*)} (nx)^{\theta^*-1} (n(1-x))^{\lambda-\theta^*-1} n^{-(\lambda-1)} \\ \sim dx \frac{\Gamma(\lambda)}{\Gamma(\theta^*)\Gamma(\lambda-\theta^*)} x^{\theta^*-1}(1-x)^{\lambda-\theta^*-1}, \end{array} \tag{13}$$

a beta density with parameters $\theta^* = \lambda/p, \lambda - \theta^* = \lambda(1-1/p)$. This shows (with $n^{-1} = dx$) that, in this asymptotic regime, $n^{-1}K(1) \xrightarrow{d} B(\theta^*, \lambda - \theta^*)$ as $n \to \infty$. This asymptotic regime was not considered in [9] either.

Note that, from (4), the limiting probability α^* that two randomly chosen individuals from the population are of the same species is

$$\alpha = \frac{p + \nu(n-1)}{p(1+\nu(n-1))} \to \alpha^* = \frac{p+\lambda}{p(1+\lambda)},$$

approaching $1/(1+\lambda)$ if p is in turn large enough.

Remark: Regime **1** deals with a large population of size n together with a large number of types p both of the same order of magnitude (a case with asymptotic density $n/p \to \mu^*$). It is balanced. In the regimes **2** to **4**, $n \ll p$, a dilute phase situation with low density of individuals compared to the species range. And while scrolling from regimes **2** to **4,** $n\nu$ ranges from infinity to zero, through moderate in regime **3**. The main results are from [9]. In the dense (large density) regime **5** on the contrary, $n \gg p$ and the population is made of few types but a large number of individuals. It is sometimes adapted to the surname distribution studies: for instance in France, there are about $p = 1.5$ million different surnames for a population of about $n = 67$ millions people. As of 2000, about $p = 286$ Korean family names were reported in use in South Korea for a population around $n = 50$ millions people. In both cases however, n cannot be assumed having stabilized. The study [22] dealing with the Sardinian island looks convincing. Note that the whole KMG theory breaks down would the hypothesis of a constant population size be relaxed, as in the Yule approach to the speciation process briefly addressed in the introduction. A hint of the drastic changes to be made in the neutral context when the population size is held constant on average only is to be found in [6]. Note also that there is no "selection effect" in the model, the adjunction of which would also considerably alter the KMG machinery ([9] p. 422).

4. Joint distributions of species abundances under KMG mutation model

So far we only obtained useful information on the limiting occupancy of a typical species $K(1)$ and only partial (mean and variance) information on the asymptotic number Q of filled species. We are able to be more precise. We start with fixed n and p before considering asymptotic regimes.

With $\theta = \frac{n}{p}\nu/(1-\nu)$, consider the Dirichlet continuous density function, say $D_p(\theta)$, on the simplex $\left\{s_q \in (0,1) : \sum_{q=1}^p s_q = 1\right\}$

$$f_{S_1,...,S_p}(s_1,...,s_p) = \frac{\Gamma(p\theta)}{\Gamma(\theta)^p}\prod_{q=1}^p s_q^{\theta-1}\cdot\delta_{\left(\sum_{q=1}^p s_q - 1\right)}. \tag{14}$$

The law of $\mathbf{S}_p := (S_1,...,S_p)$ can as well be characterized by its joint moment function ($\lambda_q > 0$)

$$\mathbf{E}\left(\prod_{q=1}^p S_q^{\lambda_q}\right) = \frac{1}{[p\theta]_{\sum_{q=1}^p \lambda_q}}\prod_{q=1}^p [\theta]_{\lambda_q}. \tag{15}$$

where $[\theta]_\lambda = \Gamma(\theta+\lambda)/\Gamma(\theta)$.

Clearly, the equilibrium joint distribution of $\mathbf{K}_t$, namely $\mathbf{K} := (K(q), q = 1,...,p)$, is a $D_p(\theta)$ $\mathbf{s}$-mixture of a multinomial multin$(n,\mathbf{s})$ distribution where $\mathbf{s} = (s_1,...,s_p)$. With $\mathbb{N}_0 := \{0,1,2,...\}$, it is thus a Dirichlet-multinomial distribution on the now discrete simplex $\left\{k_q \in \mathbb{N}_0 : \sum_{q=1}^p k_q = n\right\}$ with (see [8], Theorem 6, for instance)

$$\mathbf{P}(\mathbf{K}=\mathbf{k}) = \mathbf{E}\mathbf{P}(\mathbf{K}=\mathbf{k} \mid \mathbf{S}_p) = \frac{n!}{[p\theta]_n}\prod_{q=1}^p \frac{[\theta]_{k_q}}{k_q!}. \tag{16}$$

Here $\mathbf{K} \mid \mathbf{S}_p \overset{d}{\sim} \text{multin}(n, \mathbf{S}_p)$ and $\mathbf{S}_p \overset{d}{\sim} D_p(\theta)$. It is an exchangeable distribution, each margin being identically distributed, but of course, owing to $\sum_{q=1}^{p} K(q) = n$, the $K(q)$'s are not independent. We observe that, equivalently, with all $u_q \in (0,1)$, the joint probability generating function of $\mathbf{K}$ is

$$\mathbf{E}\left(\prod_{q=1}^{p} u_q^{K(q)}\right) = \mathbf{E}\left[\left(\sum_{q=1}^{p} u_q S_q\right)^n\right] \tag{17}$$

from which joint statistical information can be extracted using moment identities of the Dirichlet distribution. The simplest one is the (negative) covariance between any two pairs $(K(1), K(2))$ of equilibrium species abundances which can easily be found to be from (17) and using (15)

$$\text{Cov}(K(1), K(2)) = -\frac{\sigma^2(K(1))}{p-1} = -\frac{n}{p}\left(\frac{n}{p} - \frac{(n-1)\theta}{p\theta+1}\right) = -\left(\frac{n}{p}\right)^2 \frac{1}{1+\nu(n-1)}.$$

Coming back to (16), it is convenient to introduce the related joint probability

$$\mathbf{P}(K(1) = k_1, ..., K(q) = k_q; Q = q) = \binom{p}{q} \frac{n!}{[p\theta]_n} \prod_{q'=1}^{q} \frac{[\theta]_{k_{q'}}}{k_{q'}!} \tag{18}$$

where $1 \leq q \leq p$ and then with $\sum_{q'=1}^{q} k_{q'} = n$ and all $k_{q'} \geq 1$. It is the joint probability that there are $q \in \{1, ..., p\}$ **filled** species cells and that $(k_1, ..., k_q)$ are their effective abundance occupancies. Letting $\sigma_n(\theta) := n![x^n] Z_\theta(x) = [\theta]_n$ where $Z_\theta(x) = e^{\theta\phi(x)}$ and $\phi(x) = -\log(1-x)$, with $\mathbb{N} := \{1, 2, ...\}$, $\mathbf{k}_q := (k_1, ..., k_q)$ and $|\mathbf{k}_q| = \sum_{q'=1}^{q} k_{q'}$, we have

$$\mathbf{P}(Q = q) = \binom{p}{q} \frac{n!}{\sigma_n(p\theta)} \sum_{\mathbf{k}_q \in \mathbb{N}^q:\ |\mathbf{k}_q| = n} \prod_{q'=1}^{q} \frac{\sigma_{k_{q'}}(\theta)}{k_{q'}!},\ q = 1, ..., p. \tag{19}$$

The expression (18) turns out to be the canonical Gibbs distribution on the simplex $\{\mathbf{k}_q \in \mathbb{N}^q :\ |\mathbf{k}_q| = n\}$, the finite size-$p$ partitions of n into q distinct clusters (the filled species). In this language, the normalizing quantity $\sigma_n(p\theta)/n!$ is called the canonical Gibbs partition function.

Now, from (19), with $(p)_q := p!/(p-q)!$

$$\mathbf{P}(Q = q) = \frac{(p)_q}{\sigma_n(p\theta)} B_{n,q}(\sigma_\bullet(\theta)),\ q \in \{1, ..., p \wedge n\}, \tag{20}$$

where

$$B_{n,q}(\sigma_\bullet(\theta)) := \frac{n!}{q!} \sum_{\mathbf{k}_q \in \mathbb{N}^q:\ |\mathbf{k}_q| = n} \prod_{q'=1}^{q} \frac{\sigma_{k_{q'}}(\theta)}{k_{q'}!} = \frac{n!}{q!}[x^n](Z_\theta(x) - 1)^q \tag{21}$$

are the Bell polynomials in the polynomial variables $\sigma_\bullet(\theta) := (\sigma_1(\theta), \sigma_2(\theta), ...)$, [1]. Here again $Z_\theta(x) = (1-x)^{-\theta}$ and $\sigma_n(\theta) = n![z^n] e^{-\theta\log(1-z)} = [\theta]_n$.

Conditioning the canonical Gibbs distribution on the number of filled species being equal to q yields the corresponding micro-canonical distribution as

$$\mathbf{P}(K(1) = k_1, ..., K(q) = k_q \mid Q = q) = \frac{n!}{q!} \frac{1}{B_{n,q}(\sigma_\bullet(\theta))} \prod_{q'=1}^{q} \frac{\sigma_{k_{q'}}(\theta)}{k_{q'}!}. \tag{22}$$

The new normalizing constant $B_{n,q}(\sigma_\bullet(\theta))/n!$ may be called the microcanonical partition function. The special feature of the occupancy distributions (18), (20) and (22) is that $\theta = \frac{n}{p}\nu/(1-\nu)$ depends on n, p and ν.

Let us now first characterize the full distribution of Q which depends on n, p and ν (via θ) before any asymptotics is considered. We have:

(a) Assume $n \geq p$. With $u \in [0,1]$, the probability generating function of Q is given by

$$\mathbf{E}\left(u^Q\right) = \sum_{q=0}^{p-1} \binom{p}{q} u^{p-q}(1-u)^q \frac{\sigma_n((p-q)\theta)}{\sigma_n(p\theta)}, \text{ with} \tag{23}$$

$$\mathbf{P}(Q=q) = \frac{\binom{p}{q}}{\sigma_n(p\theta)} \sum_{q'=1}^{q} (-1)^{q-q'} \binom{q}{q'} \sigma_n(q'\theta), \ q \in \{1,...,p\}. \tag{24}$$

In addition,

$$\mathbf{E}(Q) = p\left(1 - \frac{\sigma_n((p-1)\theta)}{\sigma_n(p\theta)}\right) \text{ and} \tag{25}$$

$$\sigma^2(Q) = p\left(\frac{\sigma_n((p-1)\theta)}{\sigma_n(p\theta)} + (p-1)\frac{\sigma_n((p-2)\theta)}{\sigma_n(p\theta)} - p\left(\frac{\sigma_n((p-1)\theta)}{\sigma_n(p\theta)}\right)^2\right). \tag{26}$$

(b) If $n < p$, (23) and (24) still hold, but now with a modified support for Q's law:

$$\mathbf{P}(Q=q) = \frac{\binom{p}{q}}{\sigma_n(p\theta)} \sum_{q'=1}^{q} (-1)^{q-q'} \binom{q}{q'} \sigma_n(q'\theta), \ q \in \{1,...,n\}. \tag{27}$$

Statement (a) follows from $B_{n,q}(\sigma_\bullet(\theta)) = \frac{n!}{q!}[x^n](Z_\theta(x)-1)^q$. Indeed, from (20)

$$\begin{array}{l}\mathbf{E}\left(u^Q\right) = \sum_{q=0}^{p} u^q (p)_q \frac{B_{n,q}(\sigma_\bullet(\theta))}{\sigma_n(p\theta)} = \frac{n!}{\sigma_n(p\theta)} \sum_{q=0}^{p} \binom{p}{q} [x^n](u(Z_\theta(x)-1))^q \\ = \frac{n!}{\sigma_n(p\theta)}[x^n](1-u+uZ_\theta(x))^p = \frac{n!}{\sigma_n(p\theta)} \sum_{q=0}^{p} \binom{p}{q} u^{p-q}(1-u)^q [x^n] Z_\theta(x)^{p-q} \\ = \sum_{q=0}^{p-1} \binom{p}{q} u^{p-q}(1-u)^q \frac{\sigma_n((p-q)\theta)}{\sigma_n(p\theta)}.\end{array}$$

The alternating sum expression of $\mathbf{P}(Q=q)$ follows from extracting $[u^q]\mathbf{E}\left(u^Q\right)$ and the mean and variance of Q from the evaluations of the first and second derivatives of $\mathbf{E}\left(u^Q\right)$ with respect to u at $u=1$.

Statement (b) follows from similar considerations. Indeed, in principle, we should start with $\mathbf{E}\left(u^Q\right) = \sum_{q=0}^{n} u^q (p)_q \frac{B_{n,q}(\sigma_\bullet(\theta))}{\sigma_n(p\theta)}$ where the $q-$sum now stops at $q = n = p \wedge n$. But the upper bound of this $q-$sum can be extended to p because $B_{n,q}(\sigma_\bullet(\theta)) = 0$ if $q > n$.

In the particular mutation case discussed here, $\sigma_n(\theta) = [\theta]_n = \theta(\theta+1)...(\theta+n-1) = \Gamma(\theta+n)/\Gamma(\theta)$ (the Ewens-Dirichlet model, [4]). From (24), (25), for instance,

$$\mathbf{P}(Q=1) = p\frac{\sigma_n(\theta)}{\sigma_n(p\theta)} = p\frac{[\theta]_n}{[p\theta]_n} = p\frac{\Gamma(\theta+n)}{\Gamma(\theta)}\frac{\Gamma(p\theta)}{\Gamma(p\theta+n)}.$$

$$\mathbf{E}(Q) = p\left(1 - \frac{[(p-1)\theta]_n}{[p\theta]_n}\right) = p\left(1 - \frac{\Gamma((p-1)\theta+n)}{\Gamma((p-1)\theta)}\frac{\Gamma(p\theta)}{\Gamma(p\theta+n)}\right)$$

We now illustrate some of the consequences of the latter expressions under three asymptotic regimes discussed earlier.

Regime **1.** If n and $p \to \infty$ while $n/p \to \mu^*$ as in the balanced regime **1**, then $\theta = \frac{n}{p}\nu/(1-\nu) \to \theta^* = \mu^*\nu/(1-\nu)$ and, from (25), in a consistent way with previous results,

$$\begin{aligned}&\mathbf{E}(Q) \sim p\left(1 - \tfrac{\Gamma(p(\theta^*+\mu^*)-\theta^*)}{\Gamma(p(\theta^*+\mu^*))}\tfrac{\Gamma(p\theta^*)}{\Gamma(p\theta^*-\theta^*)}\right)\\ &\sim p\left(1 - \tfrac{(p(\theta^*+\mu^*))^{-\theta^*}}{(p\theta^*)^{-\theta^*}}\right) = p\left(1 - \left(\tfrac{\theta^*}{\theta^*+\mu^*}\right)^{\theta^*}\right) = p\left(1-\nu^{\theta^*}\right).\end{aligned}$$

Proceeding similarly, from (26) $\sigma^2(Q) \sim p\left(\nu^{\theta^*} - \nu^{2\theta^*}\right)$, suggesting $(Q - \mathbf{E}(Q))/\sigma(Q)$ is asymptotically normal in regime **1** as well.

From (17) and (15), with $1 \le q < p$, $a = \nu/(1-\nu)$ and $p\theta \sim na$ and $\theta \sim \theta^* = \mu^* a$

$$\begin{aligned}&\mathbf{E}\left(\textstyle\prod_{q'=1}^q u_q^{K(q')}\right) = \mathbf{E}\left[\left(1 + \sum_{q'=1}^q (u_{q'}-1) S_{q'}\right)^n\right]\\ &= \textstyle\sum_{k_1+...+k_{q+1}=n}\binom{n}{k_1...k_{q+1}}\prod_{q'=1}^q (u_{q'}-1)^{k_{q'}}\,\mathbf{E}\left(\prod_{q'=1}^q S_{q'}^{k_{q'}}\right)\\ &= \textstyle\sum_{k_1+...+k_{q+1}=n}\binom{n}{k_1...k_{q+1}}\prod_{q'=1}^q (u_{q'}-1)^{k_{q'}}\,\frac{\prod_{q'=1}^q [\theta]_{k_{q'}}}{[p\theta]_{\sum_{q'=1}^q k_{q'}}}\\ &= \textstyle\sum_{k'_{q+1}=0}^n \frac{1}{(n-k'_{q+1})!}\sum_{k_1+...+k_q=k'_{q+1}}\binom{n}{k_1...k_q}\prod_{q'=1}^q (u_{q'}-1)^{k_{q'}}\,\frac{\prod_{q'=1}^q [\theta]_{k_{q'}}}{[p\theta]_{k'_{q+1}}}\\ &\sim \textstyle\sum_{k'_{q+1}=0}^n \binom{n}{k'_{q+1}}(na)^{-k'_{q+1}}\sum_{k_1+...+k_q=k'_{q+1}}\binom{k'_{q+1}}{k_1...k_q}\prod_{q'=1}^q\left([\theta^*]_{k_{q'}}(u_{q'}-1)^{k_{q'}}\right)\\ &\sim \textstyle\sum_{k'_{q+1}=0}^\infty a^{-k'_{q+1}}\sum_{k_1+...+k_q=k'_{q+1}}\prod_{q'=1}^q\left([\theta^*]_{k_{q'}}(u_{q'}-1)^{k_{q'}}\right)/k_{q'}!,\end{aligned}$$

the pgf of the multivariate negative binomial distribution of $(K(q'); q' = 1, ..., q)$. Note that if $q = 1$, this pgf reduces, as required from Section 3, to

$$\mathbf{E}\left(u_1^{K(1)}\right) = \sum_{k=0}^\infty a^{-k}\frac{[\theta^*]_k}{k!}(u_1-1)^k = \left(\frac{\nu}{1-(1-\nu)u_1}\right)^{\theta^*}.$$

Regime **2.** (infinitely many possible types in the population). First fix population size n. If now as in regime **2**, we let $p \to \infty$ and $\beta \to 0$ (small mutation probability) while $p\beta = \nu > 0$ is fixed, then $\theta = \frac{n}{p}\nu/(1-\nu) \to 0$ while $p\theta \sim n\nu/(1-\nu) =: \gamma$.

Recall $\sigma_n(\theta) := n![x^n]Z_\theta(x) = [\theta]_n$ where $Z_\theta(x) = e^{\theta\phi(x)}$, $\phi(x) = -\log(1-x)$ and $\phi_i = \left[x^i\right]\phi(x) = (i-1)!$. We have $B_{n,q}(\phi_\bullet) = \frac{n!}{q!}[x^n]\phi(x)^q = s_{n,q}$, the first kind absolute Stirling numbers, [1].

When $\theta \to 0$, , from (21), $B_{n,q}(\sigma_\bullet(\theta)) = \frac{n!}{q!}[x^n](Z_\theta(x)-1)^q \sim \frac{n!}{q!}\theta^q[x^n]\phi(x)^q = \theta^q B_{n,q}(\phi_\bullet) = \theta^q s_{n,q}$. Thus, recalling $p\theta \sim \gamma$, (20) becomes

$$\mathbf{P}(Q=q) = \frac{(p)_q}{\sigma_n(p\theta)}B_{n,q}(\sigma_\bullet(\theta)) \sim \frac{(p\theta)^q s_{n,q}}{\sigma_n(p\theta)} = \frac{\gamma^q s_{n,q}}{\sigma_n(\gamma)},\ q = 1, ..., n, \tag{28}$$

giving the simple asymptotic shape of the law of Q for a size n population with infinitely many types. It depends on ν, via $\gamma = n\nu/(1-\nu)$. The probability generating function of this limiting Q is

$$\mathbf{E}\left(u^{Q}\right) = \frac{\sigma_n(\gamma u)}{\sigma_n(\gamma)} = \frac{[\gamma u]_n}{[\gamma]_n} = u\prod_{q=1}^{n-2}\left(\frac{\gamma u+q}{\gamma+q}\right),$$

showing that $Q \overset{d}{=} 1+\sum_{q=1}^{n-2} B_q$ where the B_q's are independent Bernoulli random variables with success parameters $\frac{\gamma}{\gamma+q}$ where $\gamma = n\nu/(1-\nu)$. Recalling $\psi(z) \underset{z\to\infty}{\sim} \log z$ and because

$$\sum_{q=0}^{n-1}\frac{\gamma}{\gamma+q} = \frac{n\nu}{1-\nu}\left(\psi\left(\frac{n}{1-\nu}\right)-\psi\left(\frac{n\nu}{1-\nu}\right)\right),$$

$\sum_{q=0}^{n-1}\frac{\gamma}{\gamma+q} \sim \frac{n\nu}{1-\nu}\log(1/\nu)$ and by strong law of large numbers $Q/n \underset{n\to\infty}{\to} \nu\log(1/\nu)/(1-\nu)$ almost surely (completing (9)).[1]

Owing now to $\sigma_{k_{q'}}(\theta) = [\theta]_{k_{q'}} \sim \theta\Gamma(k_{q'}) = \theta(k_{q'}-1)!$, with the k_q's positive summing to n, from (18),

$$\begin{aligned}&\mathbf{P}(K(1)=k_1,...,K(q)=k_q;Q=q) = \tbinom{p}{q}\tfrac{n!}{[p\theta]_n}\textstyle\prod_{q'=1}^{q}\frac{[\theta]_{k_{q'}}}{k_{q'}!}\\ &\sim \tfrac{n!}{q!}\tfrac{\gamma^q}{[\gamma]_n}\textstyle\prod_{q'=1}^{q}\frac{1}{k_{q'}!},\end{aligned}$$

and from (22)

$$\begin{aligned}&\mathbf{P}(K(1)=k_1,...,K(q)=k_q \mid Q=q) = \tfrac{n!}{q!}\tfrac{1}{B_{n,q}(\sigma_\bullet(\theta))}\textstyle\prod_{q'=1}^{q}\frac{\sigma_{k_{q'}}(\theta)}{k_{q'}!}\\ &\sim \tfrac{n!}{q!}\tfrac{1}{\theta^q s_{n,q}}\textstyle\prod_{q'=1}^{q}\frac{\theta(k_{q'}-1)!}{k_{q'}!} = \tfrac{n!}{q!}\tfrac{1}{s_{n,q}}\textstyle\prod_{q'=1}^{q}\frac{1}{k_{q'}},\end{aligned}$$

the Ewens sampling formula, [4], [18]. This gives the asymptotic shape of the joint equilibrium species abundance vector, given q of them are represented at equilibrium. A curious feature of this last distribution is that it is independent of ν.

Regime **3**. If as here $n\to\infty$ and $\nu\to 0$ while $n\nu=\lambda$, then $\gamma\sim\gamma^*=\lambda$ and we are now in the asymptotic region akin to the Chinese restaurant process. For instance

$$\sum_{q=0}^{n-1}\frac{\gamma}{\gamma+q} \overset{*}{\sim} \sum_{q=0}^{n-1}\frac{\lambda}{\lambda+q} \sim \lambda\log n \tag{29}$$

and $Q/\log n \underset{n\to\infty}{\to} \lambda$ almost surely (completing (11)). The analogy is thus with a 'chinese' restaurant with infinitely many indistinguishable tables, each of which has infinite capacity. In the table filling process, the first customer sits at some table while the next one either sits at the same table or at a different one. The process continues, with each customer choosing either to sit at an occupied table with a probability proportional to the number of customers already there or at some already unoccupied table. Under the condition of Regime 3, after step n, the

[1]This formalism resembles the one of the asymptotic number Q of filled tables in the Chinese restaurant problem with n customers (see Sections 3.1 and 3.2 of [14]). However its asymptotic behavior is of a different nature because here γ depends on n (and ν), leading to Q of order n rather than $\log n$.

occupancies of the tables are given by (18) and the n customers will be partitioned among $Q \leq n$ filled tables (or blocks of the partition) with Q of order $\log n$. The outcomes of this process are exchangeable as the order in which the customers sit does not affect the probability of the final distribution.

If, as in regime **4**, $n \to \infty$ and $\nu \to 0$ while $n\nu \sim c/\log n$ for some $c > 0$, then $\gamma \sim \gamma^* = c/\log n$ and $\mathbf{E}(Q) = \sum_{q=0}^{n-1} \frac{\gamma}{\gamma+q} \overset{*}{\sim} 1 + c$. It is easy to check that here $Q - 1 \overset{*}{\to} \text{Poi}(c)$, a Poisson random variable with mean c.

Regime **5.** (Finitely many types and very large population size). Finally, from (17), with X_q, $q = 1, ..., p$ iid Gamma(θ) random variables and $\widetilde{X}_q := X_q / \sum_{q=1}^p X_q$ and exploiting the Gamma structure of Dirichlet distributions,

$$\begin{array}{l} \mathbf{E}\left(\prod_{q=1}^p u_q^{K(q)/n}\right) = \mathbf{E}\left[\left(\sum_{q=1}^p u_q^{1/n} S_q\right)^n\right] = \frac{1}{[p\theta]_n}\mathbf{E}\left[\left(\sum_{q=1}^p u_q^{1/n} X_q\right)^n\right] \\ \underset{n\uparrow\infty}{\sim} \frac{1}{[p\theta]_n}\mathbf{E}\left[\left(\sum_{q=1}^p X_q\right)^n \left(1 + \frac{1}{n}\sum_{q=1}^p \widetilde{X}_q \log u_q\right)^n\right] \\ \underset{n\uparrow\infty}{\sim} \mathbf{E}\left(\prod_{q=1}^p u_q^{\widetilde{X}_q}\right) = \mathbf{E}\left(\prod_{q=1}^p u_q^{S_q}\right). \end{array}$$

Thus, generalizing (13),

$$\mathbf{K}/n \overset{d}{\to} \mathbf{S}_p \text{ as } n \to \infty. \tag{30}$$

Applying the strong law of large numbers (conditionally given $\mathbf{S}_p$), the above convergence in law also holds almost surely.

With the main results being from (18-27), the present study can perhaps be summarized for the different regimes as follows:

\	Range of the parameters	$Q \overset{d}{\sim}$	$\mathbf{E}(Q) \sim$	$K(1) \overset{d}{\sim}$	$\mathbf{K} \overset{d}{\sim}$
1	$n, p \to \infty$, $\frac{n}{p} = \mu^*$ ν fixed	(28)	$n\frac{1-\nu^{\theta^*}}{\mu^*}$ $\theta^* = \frac{\mu^*\nu}{1-\nu}$	negative binomial	multivariate neg. binomial
2	$n, p \to \infty$, $\frac{n}{p} \to 0$ ν fixed	(28) $\gamma = \frac{n\nu}{1-\nu}$	$n\frac{\nu \log(1/\nu)}{1-\nu}$	$\mid K(1) \geq 1$: log-series	Ewens $\gamma = \frac{n\nu}{1-\nu}$
3	$n \to \infty$, $\nu \to 0$ $\nu n = \lambda$, $\nu p \to \infty$	(28) $\gamma \sim \lambda$	$\lambda \log n$	(3)	Ewens $\gamma \sim \lambda$
4	$n \to \infty$, $\nu \to 0$ $\nu n \log n = c$	$1 + \text{Poi}(c)$	$1 + c$	(3)	(16)
5	$n \to \infty$, $\nu \to 0$ p fixed, $n\nu = \lambda$	(23)	(25)	$n\text{beta}(\theta^*, \lambda - \theta^*)$ $\theta^* = \lambda/p$	$n\mathbf{S}_p$, (30)

Acknowledgments: The author acknowledges partial support from the labex MME-DII (*Modèles Mathématiques et Économiques de la Dynamique, de l' Incertitude et des Interactions*), ANR11-LBX-0023-01. This work also benefited from the support of the Chair "*Modélisation Mathématique et Biodiversité*" of Veolia-Ecole Polytechnique-MNHN-Fondation X.

References

[1] Comtet, L. (1970). *Analyse combinatoire.* Tomes 1 et 2. Presses Universitaires de France, Paris.

[2] Engen, S. (1974). On species frequency models. Biometrika**,** Vol. 61, 263-270.

[3] Engen, S. (1978). *Stochastic abundance models.* Monographs on Applied Probability and Statistics, Chapman and Hall, London.

[4] Ewens, W.J. (1972). The sampling theory of selectively neutral alleles. Theoretical Population Biology, Vol. 3, Isue 1, pp. 82-112.

[5] Fisher, R.A., Corbet, A.S., & 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, pp. 42-58.

[6] Grosjean, N., & Huillet, T. (2017). Wright-Fisher-like models with constant population size on average. International Journal of Biomathematics 10(6), pp. 1750078.

[7] Huillet, T. (2013). Fluctuations analysis of finite discrete birth and death chains with emphasis on Moran models with mutations. ISRN Biomathematics, Vol. 2013, Article ID 939308, 21 pages.

[8] Huillet, T. (2005). Sampling formulae arising from random Dirichlet populations . Communications in Statistics - Theory and Methods, Taylor & Francis, 34 (5), pp.1019-1040.

[9] Karlin, S., & McGregor, J.L. (1967). The number of mutant forms maintained in a population. Proc. Fifth Berkeley Symp. Math. Statist. Prob, 4, pp. 415-438.

[10] Karlin, S., & McGregor, J.L. (1962). On a genetics model of Moran. Math. Proc. of the Cambridge Philos. Soc. Volume 58, Issue 2, pp. 299-311.

[11] Kimura, M., & Crow, J.F. (1964). The number of alleles that can be maintained in a finite population. Genetics, 49, pp.725-738.

[12] Manrubia, S.C., & Zanette, D.H. (2002). At the boundary between biological and cultural evolution: The origin of surname distributions. Journal of Theoretical Biology, Vol. 216, Issue 4, pp. 461-477.

[13] Moran, P.A.P. (1962). *The Statistical Processes of Evolutionary Theory.* Oxford, Clarendon Press.

[14] Pitman, J. (2006). *Combinatorial stochastic processes.* (Ecole d'Eté de Probabilités de Saint-Flour XXXII - 2002). Berlin: Springer-Verlag.

[15] Rossi, P. (2013). Surname distribution in population genetics and in statistical physics. Physics of Life Reviews, 10, pp. 395-415.

[16] Simkin, M.V., & Roychowdhury, V.P. (2011). Re-inventing Willis. Physics Reports, Vol. 502, Issue 1, pp. 1-35.

[17] Simon, H.A. (1955). On a Class of Skew Distribution Functions. Biometrika, Vol. 42, No. 3/4, pp. 425-440.

[18] Tavaré, S., & Ewens, W.J. (1997). The Multivariate Ewens distribution. Chapter 41 of : N.L. Johnson, S. Kotz, and N. Balakrishnan *Discrete Multivariate Distributions*, Wiley.

[19] Willis, J.C., & Yule, G.U. (1922). Some statistics of evolution and geographical distribution in plants and animals, and their significance, Nature, 109, 177.

[20] Yasuda, N., Cavalli-Sforza, L.L., & Skolnick, M. (1974). The evolution of surnames: an analysis of their distribution and extinction. Theoretical Population Biology, Vol. 5, Issue 1, pp 123-142.

[21] Yule, G.U. (1925). A Mathematical Theory of Evolution, Based on the Conclusions of Dr. J. C. Willis, F.R.S. Philosophical Transactions of the Royal Society of London, Series B, Vol. 213, pp. 21-87.

[22] Zei, G., Guglielmino, C.R., Siri E., Moroni A., & Cavalli-Sforza L. (1983). Surnames as neutral alleles: observations in Sardinia. Human Biology, 55, pp. 357-365.

Laboratoire de Physique Théorique et Modélisation, CNRS-UMR 8089 et Université de Cergy-Pontoise, 2 Avenue Adolphe Chauvin, 95302, Cergy-Pontoise, FRANCE, E-mail: Thierry.Huillet@u-cergy.fr