Adv. Appl. Prob. **23,** 229–258 (1991)
Printed in N. Ireland
© *Applied Probability Trust* 1991

CONSISTENT ORDERED SAMPLING DISTRIBUTIONS: CHARACTERIZATION AND CONVERGENCE

PETER DONNELLY,* *Queen Mary and Westfield College, London*
PAUL JOYCE,** *University of Southern California*

Abstract

This paper is concerned with models for sampling from populations in which there exists a total order on the collection of types, but only the relative ordering of types which actually appear in the sample is known. The need for consistency between different sample sizes limits the possible models to what are here called 'consistent ordered sampling distributions'. We give conditions under which weak convergence of population distributions implies convergence of sampling distributions and conversely those under which population convergence may be inferred from convergence of sampling distributions. A central result exhibits a collection of 'ordered sampling functions', none of which is continuous, which separates measures in a certain class. More generally, we characterize all consistent ordered sampling distributions, proving an analogue of de Finetti's theorem in this context. These results are applied to an unsolved problem in genetics where it is shown that equilibrium age-ordered population allele frequencies for a wide class of exchangeable reproductive models converge weakly, as the population size becomes large, to the so-called GEM distribution. This provides an alternative characterization which is more informative and often more convenient than Kingman's (1977) characterization in terms of the Poisson–Dirichlet distribution.

EXCHANGEABILITY; WEAK CONVERGENCE; RANDOM RANKINGS; PARTITION STRUCTURES; GEM DISTRIBUTION; AGE-ORDERED ALLELE FREQUENCIES; POISSON–DIRICHLET DISTRIBUTION; GENEALOGY

1. Introduction

Consider a (possibly infinite) population of individuals, with each of which is associated a type or label. Further suppose that the collection of possible types, $\mathcal{T}$ say, may be (totally) ordered and denote the order by $>$. Our concern is with the modelling of sampling from such a population in the case when the population itself is unobservable and the entire collection of possible types unknown, but the relative ordering of the types which occur in the sample is known. That is, if the sample contains l different types, $T_{i_1}, T_{i_2}, \cdots T_{i_l}$, say, the restriction of $>$ to the set $\{T_{i_1}, T_{i_2}, \cdots T_{i_l}\} \subset \mathcal{T}$ is known, so that the individuals in the sample may be ranked in a natural way: if the type of the ith individual in the sample is T_i and that of the

Received 4 April 1989; revision received 20 April 1990.

* Postal address: School of Mathematical Sciences, Queen Mary and Westfield College, University of London, Mile End Road, London E1 4NS, UK.

Research partially supported by NSF grant DMS 86-08857.

** Postal address: Department of Mathematics, University of Southern California, Los Angeles, CA 90089-1113, USA.

https://doi.org/10.2307/1427746 Published online by Cambridge University Press

jth, T_j, we say that i ranks above j in the sample if $T_i > T_j$, that i ranks below j if $T_j > T_i$ and that i and j share the same rank if $T_i = T_j$. Note that the same framework is appropriate if the initial assignment of types to individuals in the population is random.

Denote by B_n the set

$$\{\eta = (l; \eta_1, \eta_2, \cdots, \eta_l) : l = 1, 2, \cdots, n, \eta_i \in \mathbb{N} \text{ for each } i, \text{ and } \eta_1 + \cdots + \eta_l = n\}.$$

(Here and throughout we use $\mathbb{N}$ to denote the set $\{1, 2, 3, \cdots\}$.) Within the above context a sample of size n from the population may be described by a point $\eta = (l; \eta_1, \cdots, \eta_l) \in B_n$. The interpretation is that the sample contains l distinct types, and if we write the types in order as $T_{(1)} > T_{(2)} > \cdots > T_{(l)}$, then there are η_i representatives of the type $T_{(i)}$ in the sample, $i = 1, 2, \cdots, l$. A function $P_n : B_n \to [0, 1]$ will be said to be an *ordered sampling distribution* if

$$\sum_{\eta \in B_n} P_n(\eta) = 1.$$

There is, however, nothing special about the particular value of n under consideration and a model for this situation should define sampling distributions for a range of values of n. Furthermore, the observation that one way of obtaining a sample of size n is first to take a sample of size $m > n$ and to choose a subsample of size n from this sample forces a natural consistency on the distributions P_n for different values of n. Say that a collection $\{P_n\}_{n \in I}$ of ordered sampling distributions is *consistent* if for each $n, m \in I$, $n < m$, and each $\eta = (l; \eta_1, \cdots, \eta_l) \in B_n$

$$P_n(\eta) = \sum_{\alpha \in B_m} \binom{m}{n}^{-1} \sum_{i_1 < i_2 < \cdots < i_l \leqq k} \binom{\alpha_{i_1}}{\eta_1} \cdots \binom{\alpha_{i_l}}{\eta_l} P_m(\alpha) \tag{1}$$

where a generic element $\alpha \in B_m$ is written $\alpha = (k; \alpha_1, \cdots, \alpha_k)$. Observe that (1) is exactly the requirement that the subsample of n from m (with the ordering in the sample of n induced by that in the larger sample) should have distribution P_n when the sample of m has distribution P_m.

The motivation for this work comes from population genetics. Within that context one may associate a genetic type, or allele, with each individual. New alleles arise by mutation and the alleles present in the population or in a sample at any given time may be ordered by age: allele A_1 ranks above allele A_2 if it is older, that is if the time since it first appeared in the population (its age) is greater than the corresponding time for A_2. (The possibility that two or more alleles may share the same age is a nuisance with which we shall deal later, but for the present it will be ignored.) We note, however, that the general framework is more widely applicable. The types may correspond to species in some area in which case several different orderings may be of interest, for example a 'pecking order' or an ordering reflecting demand for a particular resource. More generally, in the spirit of recent work on measure-valued diffusions, it is possible to accommodate quite complicated and

https://doi.org/10.2307/1427746 Published online by Cambridge University Press

highly informative type spaces. If a total order on the type space does not arise naturally, one might be obtained through the amalgamation of types. Indeed, if the order is of primary interest it may not be particularly meaningful to distinguish between types which share the same ranking.

For the sake of definiteness we shall often adopt the terminology appropriate to an age ordering, and speak of the type, and sometimes of the individuals, which rank above all other types in the sample as the oldest in the sample, those ranking next as the second oldest and so on, with a similar convention for the population. It should be stressed that this is a matter of convenience. The following is much more generally applicable.

Our interest centres on two types of question. The first of these, convergence, concerns the situation in which ordered sampling distributions corresponding to a sequence of discrete populations converge to something, and asks for conditions under which this convergence of sampling distributions guarantees convergence of population distributions. Conversely it is of interest to know when population convergence implies convergence of ordered sampling distributions. These issues are discussed, and the questions answered, in Section 2. More generally, one might ask whether it is possible to characterize consistent ordered sampling distributions. In view of the structure inherent in the consistency relation (1), this corresponds (in another guise and another language) to the identification of a particular projective limit or the description of the extreme points of a particular Choquet simplex. In Section 3.1 we give a characterization result appropriate to most cases of biological relevance. Section 3.2 treats the general case by making explicit the role of exchangeability, proving an analogue of de Finetti's theorem in the context of what we call exchangeable random rankings. In the final section we use these results to solve an outstanding problem in population genetics, proving that for a wide class of reproductive models the equilibrium distribution of age-ordered allele frequencies in the *population* converges, as the population size becomes large, to a particular, tractable, distribution known as the GEM distribution. The analogous results for *samples,* which provides the starting point here, was proved in Donnelly and Tavaré (1986).

There are close connections between our work and work by Kingman (1978a, b), (1982b), see also Aldous (1985), on partition structures and exchangeable equivalence relations. In our terminology these would correspond to consistent unordered sampling distributions: there is no order relation on the types to be observed, nor is there any natural labelling of these types; the only salient feature is that it is possible to distinguish between different types. When it is appropriate, the introduction of ordering is of course more informative. We shall also see that it gives rise to a richer structure and rather more formidable technical problems (because quantities of interest no longer live naturally on the ordered simplex, a space of considerably more topological convenience than those which follow). In a similar manner, our characterization of age-ordered populations in genetics in terms of the GEM

https://doi.org/10.2307/1427746 Published online by Cambridge University Press

distribution is analogous to Kingman's (1977) characterization of the population order statistics in terms of the Poisson–Dirichlet distribution. In this setting, somewhat paradoxically, the reward for keeping track of age ordering is not only a more informative but a vastly more tractable characterization.

This is a good time to introduce some notation. A finite population of size N of which a_1 individuals are of the oldest type, a_2 of the second oldest, and so on, may be conveniently described by a point in the set

$$\Delta_N = \{x = (a_1 N^{-1}, a_2 N^{-1}, \cdots, a_N N^{-1}, 0, 0, \cdots) : a_i \in \mathbb{N} \cup \{0\},$$
$$a_1 + \cdots + a_N = N,\ a_j = 0 \Rightarrow a_k = 0 \text{ for } k = j+1, j+2, \cdots, N\}$$

which lists, in order, the proportions of each of the types present. Hypothetically infinite populations will arise as limits of finite populations and also as objects from which one might sample in order to generate consistent ordered sampling distributions. In many circumstances it is natural to describe such a population by a point in the set

$$\Delta = \left\{x = (x_1, x_2, x_3, \cdots) : x_i \geqq 0 \text{ for each } i,\ \sum_{i=1}^{\infty} x_i = 1\right\}.$$

The interpretation of this description is that a proportion x_1 of the population are of the oldest type, x_2 of the second oldest, and so on. There are two assumptions implicit in such a description. The first is that the total proportion of individuals whose types are represented only a finite number of times in the population is zero, and the second is that there is a well defined oldest, second-oldest, $\cdots$, type. Kingman (1978a, b) argues convincingly that the first assumption is appropriate in most (or all) situations of biological relevance and we would claim that the second is equally reasonable. (Of course if there is a well-defined youngest, second youngest, $\cdots$, type, the ordering could be reversed.) The point is that while more general constructions may give rise to consistent ordered sampling distributions, and herein lies the interest in the characterization results of Section 3.2, they are unlikely to arise as limits of real populations.

It is topologically convenient to regard Δ as a subset of the set

$$\bar{\Delta} = \left\{x = (x_1, x_2, \cdots) : x_i \geqq 0,\ i = 1, 2, \cdots, \sum_{i=1}^{\infty} x_i \leqq 1\right\}.$$

Topologize $\bar{\Delta}$ as a closed and hence compact subset of the product space $[0, 1]^{\mathbb{N}}$, and endow it with the Borel σ-algebra. Probability measures will arise as descriptions of random populations and are to be thought of as being defined on $\bar{\Delta}$. (The sets Δ_N and Δ defined above and Δ^* to be defined below, are easily checked to be measurable.) Throughout, the symbol $\Rightarrow$ is used to denote weak convergence.

Finally we introduce another subset of $\bar{\Delta}$ which is destined to play a central role. Let

$$\Delta^* = \{x = (x_1, x_2, \cdots) \in \bar{\Delta} : x_j = 0 \Rightarrow x_k = 0,\ k = j+1, j+2, \cdots\}.$$

https://doi.org/10.2307/1427746 Published online by Cambridge University Press

The set Δ^* consists of those points in $\bar{\Delta}$ which do not have 'zeros in the middle'. Its importance is in a sense technical, but note that *in our framework* (because of relative rather than absolute ordering) a population which has, say, no individuals of the second- or third-oldest type but some of the fourth-oldest, and hence a description $(x_1, 0, 0, x_4, \cdots) \notin \Delta^*$ (where $x_4 > 0$) is not very meaningful.

Before continuing, we remark that it may also appear natural to identify Δ with $\mathcal{P}(\mathbb{N})$ the set of probability measures on $\mathbb{N}$, and to endow it with the topology of weak convergence. Although we do not give the details, this approach is equivalent to the one we adopt here. (For example it is straightforward to verify that if $x, x^{(1)}, x^{(2)}, \cdots, \in \Delta$, then $x^{(n)} \to x$ in the product topology if and only if $x^{(n)} \Rightarrow x$ as probability measures on $\mathbb{N}$.)

2. Convergence

One way of ensuring consistency in sampling distributions is to construct them as samples from a real population. For $\eta \in B_n$ define $P_n^N(\eta, x)$ to be the probability of obtaining the configuration η when taking a sample of size n, without replacement, from a population of size N ($\geqq n$) with a description x. That is, for $x \in \Delta_N$

$$P_n^N(\eta, x) = \binom{N}{n}^{-1} \sum_{i_1 < i_2 < \cdots < i_l \leqq N} \binom{Nx_{i_1}}{\eta_1} \cdots \binom{Nx_{i_l}}{\eta_l}. \tag{2}$$

More generally, the population configuration might be random, so if μ_N is a measure on $\bar{\Delta}$ with $\mu_N(\Delta_N) = 1$, define

$$P_n^N(\eta, \mu_N) = \int_{\bar{\Delta}} P_n^N(\eta, x) \mu_N(dx).$$

It is easy to check that the family $\{P_n^N\}_{n=1}^N$ is consistent.

Alternatively, one could sample from a hypothetically infinite population. If the population description is $x \in \Delta$, then for $\eta \in B_n$,

$$\phi_\eta(x) = \sum_{i_1 < i_2 < \cdots < i_l} \frac{n!}{\eta_1! \cdots \eta_l!} x_{i_1}^{\eta_1} \cdots x_{i_l}^{\eta_l} \tag{3}$$

is the probability of obtaining a sample with configuration η. Again, if x is random with distribution μ on $\bar{\Delta}$, with $\mu(\Delta) = 1$, this probability is

$$\int_{\bar{\Delta}} \phi_\eta \, d\mu. \tag{4}$$

It is evident that the prescription (4) gives rise to a consistent family of ordered sampling distributions with index $n \in \mathbb{N}$.

Our first result is that if the finite populations converge in distribution to an infinite population described (almost surely) by a point in Δ, then the corresponding

https://doi.org/10.2307/1427746 Published online by Cambridge University Press

sampling distributions converge to (4). Note that this result is not immediate because *none* of the sampling function ϕ_η is continuous (see Remark 2 after Theorem 3). However, since $\phi_\eta = \sup_j \phi^j_\eta$, where for each j

$$\phi^j_\eta(x) \equiv \sum_{i_1<i_2<\cdots<i_l\leqq j} \frac{n!}{\eta_1!\cdots\eta_l!} x_{i_1}^{\eta_1}\cdots x_{i_l}^{\eta_l}$$

is a function depending continuously on a finite number of coordinates and hence continuous, the ϕ_η are all lower semicontinuous.

Theorem 1. *If* $\mu_N \Rightarrow \mu$ *and* $\mu(\Delta) = 1$, *then as* $N \to \infty$

$$P_n^N(\eta, \mu_N) \to \int_{\bar{\Delta}} \phi_n \, d\mu.$$

Proof. The proof is similar in spirit to parts of Kingman's (1978a) frequency limit theorem. First observe from (2) that

$$\begin{aligned} P_n^N(\eta, \mu_N) &\leqq \frac{n!}{\eta_1!\cdots\eta_l!} \int_{\bar{\Delta}} \sum_{i_1<i_2<\cdots<i_l} \frac{(Nx_{i_1})^{\eta_1}\cdots(Nx_{i_l})^{\eta_l}}{N^n - \binom{n}{2}N^{n-1}} d\mu_N. \\ &= \left(1 - \binom{n}{2}N^{-1}\right)^{-1} \int_{\bar{\Delta}} \phi_\eta \, d\mu_N. \end{aligned} \tag{5}$$

Note that for all $x \in \Delta$

$$\begin{aligned} \sum_{\eta\in B_n} \phi_\eta(x) &= \sum_{\eta\in B_n} \sum_{i_1<\cdots<i_l} \frac{n!}{\eta_1!\cdots\eta_l!} x_{i_1}^{\eta_1}\cdots x_{i_l}^{\eta_l} \\ &= \left(\sum_{i=1}^{\infty} x_i\right)^n \\ &= 1. \end{aligned} \tag{6}$$

Also

$$\sum_{\eta\in B_n} P_n^N(\eta, \mu_N) = 1. \tag{7}$$

From (5), (6), and (7), we obtain

$$\left(1 - \binom{n}{2}N^{-1}\right)(1 - P_n^N(\eta, \mu_N)) \leqq 1 - \int_{\bar{\Delta}} \phi_\eta \, d\mu_N,$$

and so

$$\left(1 - \binom{n}{2}N^{-1}\right)P_n^N(\eta, \mu_N) \leqq \int_{\bar{\Delta}} \phi_\eta \, d\mu_N \leqq \left(1 - \binom{n}{2}N^{-1}\right)P_n^N(\eta, \mu_N) + \binom{n}{2}N^{-1}. \tag{8}$$

The lower semicontinuity of ϕ_η and the hypothesis $\mu_N \Rightarrow \mu$ imply that

$$\liminf_{N\to\infty} \int_{\bar{\Delta}} \phi_\eta \, d\mu_N \geqq \int_{\bar{\Delta}} \phi_\eta \, d\mu,$$

https://doi.org/10.2307/1427746 Published online by Cambridge University Press

but using (6)

$$\limsup_{N\to\infty} \int_{\bar{\Delta}} \phi_\eta \, d\mu_N = \limsup_{N\to\infty} \left(1 - \sum_{\substack{\eta' \in B_N \\ \eta' \neq \eta}} \int_{\bar{\Delta}} \phi_{\eta'} \, d\mu_N\right)$$

$$= 1 - \sum_{\substack{\eta' \in B_N \\ \eta' \neq \eta}} \liminf_{N\to\infty} \int_{\bar{\Delta}} \phi_{\eta'} \, d\mu_N$$

$$\leqq 1 - \sum_{\substack{\eta' \in B_N \\ \eta' \neq \eta}} \int_{\bar{\Delta}} \phi_\eta \, d\mu.$$

Thus $\int_{\bar{\Delta}} \phi_\eta \, d\mu_N \to \int_{\bar{\Delta}} \phi_\eta \, d\mu$, and from (8), $P_n^N(\eta, \mu_N) \to \int_{\bar{\Delta}} \phi_\eta \, d\mu$ as required.

With any $\eta = (l; \eta_1, \cdots, \eta_l) \in B_n$ we can associate a point $x_\eta \in \Delta_n$ in a natural way: $x_\eta \equiv (n^{-1}\eta_1, n^{-1}\eta_2, \cdots, n^{-1}\eta_l, 0, 0, \cdots)$. Any sampling distribution P_n then naturally induces an atomic measure, which we call $\hat{P}_n$, on $\bar{\Delta}$: for any Borel set A, put

$$\hat{P}_n(A) = \sum_{\eta \in B_n : x_\eta \in A} P_n(\eta). \tag{9}$$

Clearly, $\hat{P}_n(\Delta_n) = 1$. If, for each n, P_n is defined by (4), for some μ with $\mu(\Delta) = 1$, this construction associates a sequence of measures $\{\hat{P}_n\}$ with μ. We now extend this construction. Let

$$A_n = \{y = (n^{-1}a_1, n^{-1}a_2, \cdots, n^{-1}a_l, 0, 0, \cdots) \in \bar{\Delta} : l \leqq n, \, a_i \in \mathbb{N} \text{ for each } i \text{ and } a_1 + \cdots + a_l \leqq n\}.$$

For $y = (n^{-1}a_1, n^{-1}a_2, \cdots, n^{-1}a_l, 0, 0, \cdots) \in A_n$, define the function $\phi_y^n : \bar{\Delta} \to [0, 1]$ by

$$\phi_y^n(x) = c_y \left(1 - \sum_{i=1}^{\infty} x_i\right)^{(n - a_1 - a_2 - \cdots - a_l)} \sum_{i_1 < \cdots < i_l} x_{i_1}^{a_1} x_{i_2}^{a_2} \cdots x_{i_l}^{a_l} \tag{10}$$

where

$$c_y = \frac{n!}{a_1! \, a_2! \cdots a_l! \, (n - a_1 - a_2 - \cdots - a_l)!}. \tag{11}$$

If μ is a measure on $\bar{\Delta}$, define the sampling measure $\hat{P}_n$ associated with μ by

$$\hat{P}_n(A) = \sum_{y \in A \cap A_n} \int_{\bar{\Delta}} \phi_y^n \, d\mu \tag{12}$$

for $n = 1, 2, \cdots$ and any Borel set $A \subseteq \bar{\Delta}$. Note that $\hat{P}_n$ is atomic and that if $\eta \in B_n$, $\phi_{y_\eta}^n(x) = \phi_\eta(x)$ for all $x \in \Delta$. Thus (12) is an extension of (4) (with P_n defined by (9)) and our use of a very similar notation for the functions (3) and (10) might be excused. Our next result says that these functions separate measures which are concentrated on Δ^*.

https://doi.org/10.2307/1427746 Published online by Cambridge University Press

Theorem 2. (a) *Suppose* μ *is a measure on* $\bar{\Delta}$ *with* $\mu(\Delta^*) = 1$. *If for each* n *and Borel set* $A \subseteq \bar{\Delta}$, *the measure* $\hat{P}_n$ *is given by*

$$\hat{P}_n(A) = \sum_{y \in A \cap A_n} \int_{\bar{\Delta}} \phi_y^n \, d\mu,$$

then as $n \to \infty$, $\hat{P}_n \Rightarrow \mu$.

(b) *Suppose* μ *and* ν *are measures on* $\bar{\Delta}$ *with* $\mu(\Delta^*) = \nu(\Delta^*) = 1$. *If for each* n *and each* $y \in A_n$,

$$\int_{\bar{\Delta}} \phi_y^n \, d\mu = \int_{\bar{\Delta}} \phi_y^n \, d\nu,$$

then $\mu = \nu$.

Proof. First note that (b) follows easily from (a). For Borel sets A, define

$$\hat{P}_n(A) = \sum_{y \in A \cap A_n} \int_{\bar{\Delta}} \phi_y^n \, d\mu = \sum_{y \in A \cap A_n} \int_{\bar{\Delta}} \phi_y^n \, d\nu$$

and use (a) to deduce that $\hat{P}_n \Rightarrow \mu$ and $\hat{P}_n \Rightarrow \nu$, so $\mu = \nu$.

Let $\mathscr{A}$ be the collection of continuous functions $f : \bar{\Delta} \to \mathbb{R}$ for which $f(x)$ depends only on the first k coordinates of x, for some $k \geqq 1$, and $0 \leqq f(x) \leqq 1$ for all $x \in \bar{\Delta}$. As a consequence of the compactness of $\bar{\Delta}$ and the Stone–Weierstrass theorem, part (a) of the theorem will follow if

$$\int_{\bar{\Delta}} f \, d\hat{P}_n \to \int_{\bar{\Delta}} f \, d\mu \tag{13}$$

for all $f \in \mathscr{A}$.

Fix $f \in \mathscr{A}$. Then, by assumption,

$$\begin{aligned} \int_{\bar{\Delta}} f \, d\hat{P}_n &= \sum_{y \in A_n} f(y) \hat{P}_n(\{y\}) \\ &= \sum_{y \in A_n} f(y) \int_{\bar{\Delta}} \phi_y^n(x) \mu(dx) \\ &= \int_{\bar{\Delta}} \sum_{y \in A_n} f(y) \phi_y^n(x) \mu(dx) \\ &= \int_{\bar{\Delta}} F_n(x) \mu(dx) \end{aligned} \tag{14}$$

say, where $F_n : \bar{\Delta} \to [0, 1]$ is given by

$$F_n(x) = \sum_{y \in A_n} \sum_{i_1 < \cdots < i_l} c_y f(y) x_{i_1}^{ny_1} \cdots x_{i_l}^{ny_l} \left(1 - \sum_{i=1}^{\infty} x_i\right)^{n(1-y_1-y_2-\cdots-y_l)}. \tag{15}$$

We show that for $x \in \Delta^*$, $F_n(x) \to f(x)$ as $n \to \infty$. Since, for some k, f only depends on the first k coordinates of x we shall often write $f(x_1, \cdots, x_k)$ for $f(x)$. The key to

https://doi.org/10.2307/1427746 Published online by Cambridge University Press

the proof is to make explicit the multinomial sampling implicit in (14) (or more fundamentally in (10)) and then to use the strong law and the structure of Δ^*. Fix $x \in \Delta^*$ and let $\{X_i\}$ be a sequence of independent and identically distributed random variables with

$$P(X_i = 0) = 1 - \sum_{j=1}^{\infty} x_j$$

and for $j = 1, 2, 3, \cdots$, $P(X_i = j) = x_j$. Define $V_j^n = |\{i : X_i = j,\ i = 1, 2, \cdots, n\}|$, $j = 1, 2, 3, \cdots$ (where $|B|$ denotes the cardinality of a set B) and for $j = 1, 2, 3, \cdots$, put $Y_j^n = n^{-1} V_j^n$. Note that the strong law of large numbers ensures that for each j, $Y_j^n \to x_j$ as $n \to \infty$. Now, for $Y^n = (Y_1^n, Y_2^n, \cdots)$ define $\sigma_k Y^n$ to be the sequence $(Y_{i_1}^n, Y_{i_2}^n, \cdots, Y_{i_k}^n)$ where $i_1, i_2, \cdots, i_k$ are defined by

$$i_1 = \inf\{i : Y_i^n > 0\}, \qquad i_2 = \inf\{i > i_1 : Y_i^n > 0\},$$

and so on. Put $\inf(\phi) = \infty$ and if $i_l = \infty$ for some l we put

$$\begin{aligned} \sigma_k Y^n &= (Y_{i_1}^n, Y_{i_2}^n, \cdots, Y_{i_{l-1}}^n, 0, 0, \cdots, 0) \\ &= (Y_{i_1}^n, Y_{i_2}^n, \cdots, Y_{i_{l-1}}^n, Y_{i_{l-1}+1}^n, Y_{i_{l-1}+2}^n, \cdots, Y_{i_{l-1}+k-l+1}^n). \end{aligned} \tag{16}$$

Note that the mapping σ_k is one way of formalizing the idea that absolute population ordering is lost, but relative ordering is maintained in the sample. With this set-up we can rewrite (15) as $F_n(x) = E(f(\sigma_k Y^n))$. The rest of the proof rests on the fact that for large enough n, since $x \in \Delta^*$, $\sigma_k Y^n$ will be $(Y_1^n, Y_2^n, \cdots, Y_k^n)$ with high probability, and the strong law, which guarantees that $(Y_1^n, Y_2^n, \cdots, Y_k^n) \to (x_1, \cdots, x_k)$ as $n \to \infty$. Denote by $C_{n,k}$ the event $\{\sigma_k Y_n = (Y_1, \cdots, Y_k)\}$, and by $C'_{n,k}$ its complement. In the light of (16),

$$C'_{n,k} = \{\exists i, j; i = 1, 2, \cdots, k, j > i, \text{ with } Y_i^n = 0,\ Y_j^n > 0\},$$

so

$$P(C'_{n,k}) \leqq \sum_{i=1}^{k} (1 - x_i)^n \left(1 - \left(1 - \sum_{j=i+1}^{\infty} x_j\right)^n\right).$$

If $x_i = 0$, then since $x \in \Delta^*$, $x_j = 0\ \forall j > i$, so the ith term in the above summation is zero. If $x_i > 0$, then the ith term in the summation converges to zero as $n \to \infty$.

Writing $1_{C_{n,k}}$ for the indicator of the event $C_{n,k}$ we have

$$E(1_{C_{n,k}} f(\sigma_k Y^n)) \leqq F_n(x) \leqq E(f(\sigma_k Y^n)).$$

As a consequence of the definition of $C_{n,k}$ and the fact that $0 \leqq f(x) \leqq 1$ for all $x \in \bar{\Delta}$, it follows that

$$E(1_{C_{n,k}} f(Y_1, \cdots, Y_k)) \leqq F_n(x) \leqq E(1_{C_{n,k}} f(Y_1, \cdots, Y_k)) + P(C'_{n,k}). \tag{17}$$

We have just shown that $P(C'_{n,k}) \to 0$ as $n \to \infty$. Thus, the strong law, and continuity of f imply that $1_{C_{n,k}} f(Y_1, \cdots, Y_k) \to f(x_1, \cdots, x_k)$ almost surely as $n \to \infty$, so dominated convergence in (17) gives $F_n(x) \to f(x_1, \cdots, x_k) = f(x)$ as $n \to \infty$. A

https://doi.org/10.2307/1427746 Published online by Cambridge University Press

further application of dominated convergence ($|F_n(x)| \leqq 1$ for all n) and the fact that $\mu(\Delta^*) = 1$ completes the proof of the theorem, since from (14) we now have

$$\int_{\bar{\Delta}} f\, d\hat{P}_n = \int_{\bar{\Delta}} F_n(x)\mu(dx) = \int_{\Delta^*} F_n(x)\mu(dx) \to \int_{\Delta^*} f(x)\mu(dx) = \int_{\bar{\Delta}} f\, d\mu.$$

We are now in a position to prove the major result of this section, a partial converse to Theorem 2 which gives conditions under which convergence of sampling distributions implies convergence of population measures.

Theorem 3. *Suppose* $\{\mu_N\}$ *is a sequence of measures on* $\bar{\Delta}$ *with* $\mu_N(\Delta_N) = 1$ *for each N. If*

$$\lim_{N\to\infty} P_n^N(\eta, \mu_N) = P_n(\eta) \text{ say},$$

exists for each n and each $\eta \in B_n$, *then* $\{P_n\}_{n\in\mathbb{N}}$ *is a consistent family of ordered sampling distributions. If for each limit point* ν *of the sequence* $\{\mu_N\}$, $\nu(\Delta^*) = 1$ *and* $\nu(\Delta) = 1$, *then there exists a* μ *with* $\mu(\Delta^*) = 1$ *such that* $\mu_N \Rightarrow \mu$ *and* $P_n(\eta) = \int_{\bar{\Delta}} \phi_\eta\, d\mu$, *for each n and each* $\eta \in B_n$.

Proof. First observe that for each n

$$\sum_{\eta\in B_n} P_n(\eta) = \sum_{\eta\in B_n} \lim_{N\to\infty} P_n^N(\eta, \mu_N) = \lim_{N\to\infty} \sum_{\eta\in B_n} P_n^N(\eta, \mu_N) = 1,$$

and for $\eta \in B_n$ and $m > n$,

$$\begin{aligned} P_n(\eta) &= \lim_{\substack{N\to\infty \\ N>m}} P_n^N(\eta, \mu_N) = \lim_{N\to\infty} \sum_{\alpha\in B_m} \sum_{i_1<i_2<\cdots<i_l\leqq k} \binom{m}{n}^{-1} \binom{\alpha_{i_1}}{\eta_1} \cdots \binom{\alpha_{i_l}}{\eta_l} P_m^N(\alpha, \mu_N) \\ &= \sum_{\alpha\in B_m} \sum_{i_1<i_2<\cdots<i_l\leqq k} \binom{m}{n}^{-1} \binom{\alpha_{i_1}}{\eta_1} \cdots \binom{\alpha_{i_l}}{\eta_l} P_m(\alpha) \end{aligned}$$

since the summation contains only a finite number of terms. (We have again written $\alpha = (k; \alpha_1, \cdots, \alpha_k)$ for a generic element of B_m.) This proves that $\{P_n\}_{n\in\mathbb{N}}$ is a consistent family of ordered sampling distributions.

Since $\bar{\Delta}$ is compact, the sequence $\{\mu_N\}$ is tight. Suppose that ν and ν' are the limits of two different weakly convergent subsequences $\{N_k\}$ and $\{N_{k'}\}$ say. By assumption $\nu(\Delta) = \nu'(\Delta) = 1$, so an application of Theorem 1 gives

$$\begin{aligned} P_n(\eta) &= \lim_{k\to\infty} P_n^{N_k}(\eta, \mu_{N_k}) = \int_{\bar{\Delta}} \phi_\eta\, d\nu, \\ P_n(\eta) &= \lim_{k\to\infty} P_n^{N_{k'}}(\eta, \mu_{N_{k'}}) = \int_{\bar{\Delta}} \phi_\eta\, d\nu' \end{aligned} \tag{18}$$

for each n and $\eta \in B_n$. We have also assumed that $\nu(\Delta^*) = \nu'(\Delta^*) = 1$, so Theorem 2(b) implies that $\nu = \nu'$. Thus all weakly convergent subsequences of $\{\mu_N\}$ converge to the same limit. Call this limit μ. Then $\mu_N \Rightarrow \mu$, $\mu(\Delta^*) = 1$ by assumption, and

https://doi.org/10.2307/1427746 Published online by Cambridge University Press

from (18), $P_n(\eta) = \int_{\bar{\Delta}} \phi_\eta \, d\mu$ for each n and each $\eta \in B_n$, so we have established the result.

The next result provides alternative conditions in the case in which the structure of the limiting consistent ordered sampling distribution is known.

Theorem 4. *Suppose* $\{\mu_N\}$ *is a sequence of measures on* $\bar{\Delta}$ *with* $\mu_N(\Delta_N) = 1$ *for each* N. *If* $\lim_{N\to\infty} P_n^N(\eta, \mu_N) = P_n(\eta)$ *say, exists for each* n *and each* $\eta \in B_n$, *then* $\{P_n\}_{n\in\mathbb{N}}$ *is a consistent family of ordered sampling distributions. Suppose there exists a* μ *with* $\mu(\Delta) = 1$ *and* $P_n(\eta) = \int_{\bar{\Delta}} \phi_\eta \, d\mu$ *for each* n *and each* $\eta \in B_n$. *If for each limit point* ν *of the sequence* $\{\mu_N\}$, $\nu(\{x : x_i > 0 \text{ for each } i\}) = 1$, *then* $\mu_N \Rightarrow \mu$.

Proof. We shall show that the given conditions imply that $\nu(\Delta) = 1$ for every limit point ν of the sequence $\{\mu_N\}$. The result will then follow from Theorem 3. First note that it is straightforward to prove that for any measure μ on $\bar{\Delta}$,

$$\mu(\Delta) = 1 \text{ if and only if } \lim_{k\to\infty} \int_{\bar{\Delta}} \sum_{i=1}^{k} x_i \mu(dx) = 1. \tag{19}$$

Let $E_{n,k}^N$ denote the event that an individual selected at random from a sample of n taken from the population described by μ_N belongs to one of the k oldest types in the sample. Then

$$\begin{aligned} P(E_{n,k}^N) &= P(\text{individual belongs to } k \text{ oldest types in population}) \\ &\quad + P(\text{at least one of } k \text{ oldest population types is not represented in sample} \cap E_{n,k}^N) \\ &\leqq P(\text{individual belongs to } k \text{ oldest types in population}) \\ &\quad + P(\text{at least one of the } k \text{ oldest population types is not represented in sample}). \end{aligned} \tag{20}$$

A calculation similar to that in the proof of Theorem 1 shows that the second probability on the right of (20) is bounded above by

$$\left(1 - \binom{n}{2} N^{-1}\right)^{-1} \int_{\bar{\Delta}} \sum_{i=1}^{k} (1 - x_i)^n \mu_N(dx).$$

The first equals $\int_{\bar{\Delta}} \sum_{i=1}^{k} x_i \mu_M(dx)$, so

$$P(E_{n,k}^N) \leqq \int_{\bar{\Delta}} \sum_{i=1}^{k} x_i \mu_N(dx) + \left(1 - \binom{n}{2} N^{-1}\right)^{-1} \int_{\bar{\Delta}} \sum_{i=1}^{k} (1 - x_i)^n \mu_N(dx). \tag{21}$$

Also

$$\begin{aligned} P(E_{n,k}^N) &= \sum_{\eta\in B_n} n^{-1}(\eta_1 + \cdots + \eta_k) P_n^N(\eta, \mu_N) \\ &\to \sum_{\eta\in B_n} n^{-1}(\eta_1 + \cdots + \eta_k) P_n(\eta) \\ &= \int_{\bar{\Delta}} \sum_{i=1}^{k} x_i \hat{P}_n(dx) \end{aligned} \tag{22}$$

https://doi.org/10.2307/1427746 Published online by Cambridge University Press

as $N \to \infty$, by assumption. (If $\eta = (l; \eta_1, \cdots, \eta_l)$ has fewer than k components, put $\eta_{l+1}, \cdots, \eta_k$ equal to zero.)

Suppose ν is the limit of some subsequence of $\{\mu_N\}$ and for convenience index the subsequence by N. Then, for any k, since $\sum_{i=1}^{k} x_i$ is continuous and bounded,

$$\begin{aligned}\int_{\bar{\Delta}} \sum_{i=1}^{k} x_i \nu(dx) &= \lim_{N\to\infty} \int_{\bar{\Delta}} \sum_{i=1}^{k} x_i \mu_N(dx) \\ &\geqq \lim_{N\to\infty} P(E^N_{n,k}) - \lim_{N\to\infty} \left(1 - \binom{n}{2} N^{-1}\right)^{-1} \int_{\bar{\Delta}} \sum_{i=1}^{k} (1-x_i)^n \mu_N(dx) \qquad (23) \\ &= \int_{\bar{\Delta}} \sum_{i=1}^{k} x_i \hat{P}_n(dx) - \int_{\bar{\Delta}} \sum_{i=1}^{k} (1-x_i)^n \nu(dx),\end{aligned}$$

in the light of (21) and (22). Now let $n \to \infty$. Theorem 2(a) ensures that $\hat{P}_n \Rightarrow \mu$, and the assumption that $\nu(\Delta^*) = 1$ and dominated convergence ensure that the second term in (23) converges to zero. Thus

$$\int_{\bar{\Delta}} \sum_{i=1}^{k} x_i \nu(dx) \geqq \int_{\bar{\Delta}} \sum_{i=1}^{k} x_i \mu(dx) \to 1$$

as $k \to \infty$, since by assumption $\mu(\Delta) = 1$. It follows from (19) that $\nu(\Delta) = 1$, and the proof is complete.

Remark 1. Since $\phi_\eta(x_1, x_2, \cdots) = \phi_\eta(0, x_1, x_2, \cdots) = \phi_\eta(x_1, 0, x_2, \cdots)$ etc., (and similarly for ϕ^n_y) it is clear that the 'avoidance of zeros' implicit in the conditions of Theorems 2, 3, and 4 is necessary.

Remark 2. If $\eta = (l; \eta_1, \cdots, \eta_l) \in B_n$ and for each k, $x^{(k)} \in \Delta$ is given by

$$x^{(k)} = \left(\underbrace{0, 0, \cdots, 0}_{k \text{ zeros}}, \frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n}, 0, 0, \cdots\right)$$

then $x^{(k)} \to (0, 0, \cdots)$ as $k \to \infty$, but $\phi_\eta(x^{(k)}) = n^{-l} c_\eta \not\to \phi_\eta((0, 0, \cdots)) = 0$. One novelty of the above results is that a collection of functions, none of which is continuous, is used to separate measures within a particular class and determine convergence within that class.

Remark 3. Lower semicontinuity and the observation that for $x \in \Delta$

$$\phi_\eta(x) = 1 - \sum_{\substack{\eta' \in B_n \\ \eta' \neq \eta}} \phi_{\eta'}(x)$$

means that as functions on Δ, the ϕ_η *are* continuous. This seems of little value, however, since a Stone–Weierstrass approximation argument requires a compact space and furthermore the ϕ_η are not continuous on Δ^*. (These comments apply equally to the ϕ^n_η).

https://doi.org/10.2307/1427746 Published online by Cambridge University Press

Remark 4. The condition $\mu_N(\Delta_N) = 1$ in both Theorems 3 and 4 merely ensures that the functions $P_n^N(\eta, \mu_N)$ defined via (2) make sense. While this requirement seems natural for the applications we envisage, it is in no other sense essential to the proof. It is clear that provided the sampling functions are defined appropriately the results will hold in greater generality, but we omit the details. (For example the measure μ_N may describe a finite population whose size is fixed but different from N or which may, as in Kingman (1978a), even be random.)

3. Characterization

3.1. *The representable case.* Throughout this section $\{P_n\}_{n\in\mathbb{N}}$ will denote a consistent family of ordered sampling distributions, and we focus attention on characterizing such structures. The main result of this subsection, Theorem 6, is that all such consistent families of biological relevance have the form (4).

Theorem 5. *If $\{P_n\}_{n\in\mathbb{N}}$ is a consistent family of ordered sampling distributions and if for each n, $\hat{P}_n$ is the measure on $\bar{\Delta}$ associated (via (9)) with P_n, then there is a measure μ on $\bar{\Delta}$ with $\mu(\Delta^*) = 1$ such that $\hat{P}_n \Rightarrow \mu$ as $n \to \infty$.*

Proof. The first part, convergence, is very similar to the proof in Kingman (1978b), so we merely sketch the details. With $\Omega = \prod_{n=1}^{\infty} \bar{\Delta}$ denoting the cartesian product of copies of $\bar{\Delta}$ (compact in the product topology) and $\mathcal{F}$ the product (and Borel) σ-algebra, the Daniel–Kolmogorov theorem and consistency guarantee the existence of a probability measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$ such that, for all n

$$\mathbb{P}\{\omega : \omega_r = x^{(r)} (r = 1, 2, \cdots, n)\} = \hat{P}_n(x^{(n)}) \prod_{r=2}^{n} S_r(x^{(r)}, x^{(r-1)})$$

where $w_r : \Omega \to \bar{\Delta}$ is the rth coordinate function and the (sampling) function S_r is defined by $S_r(x^{(r)}, x^{(r-1)}) = a_i/r$ if

$$x^{(r)} = (r^{-1}a_1, r^{-1}a_2, \cdots, r^{-1}a_l, 0, 0, \cdots) \qquad a_i \in \mathbb{N}, \qquad a_1 + \cdots + a_l = r$$
$$x^{(r-1)} = ((r-1)^{-1}a_1, \cdots, (r-1)^{-1}(a_i - 1), \cdots, (r-1)^{-1}a_l, 0, 0, \cdots)$$
$$\text{for some } i,\ i = 1, 2, \cdots, l,$$

and $S_r(x^{(r)}, x^{(r-1)}) = 0$ otherwise. Thus if $x^{(r)} \in \Delta_r$ describes a sample of size r, $S_r(x^{(r)}, x^{(r-1)})$ is the probability that when one individual is discarded at random the resulting sample of size $r-1$ will have description $x^{(r-1)}$.

For $x \in \bar{\Delta}$, define, for $k = 1, 2, \cdots$, $\alpha_k(x) = \sum_{i=1}^{k} x_i$. Thus if ω is a random element of $(\Omega, \mathcal{F}, \mathbb{P})$, we may think of $\alpha_k(\omega_r)$ as the proportion of individuals in a simple of size r (with distribution P_r) who are of the k oldest types in the sample. If the sample of r is a subsample of a larger collection of m individuals then observe that the expected proportion of the k oldest types in the smaller sample is not less than the proportion of the k oldest types in the larger sample. Hence for $m > r$

$$\mathbb{E}(\alpha_k(\omega_r) \mid \omega_m, \omega_{m+1}, \cdots) \geqq \alpha_k(\omega_m),$$

https://doi.org/10.2307/1427746 Published online by Cambridge University Press

(where $\mathbb{E}$ denotes expectation with respect to $\mathbb{P}$), so that for each $k \geqq 1$ the random sequence $\{\alpha_k(\omega_r), r = 1, 2, \cdots\}$ is a reversed submartingale with values in $[0, 1]$. Thus $y_k = \lim_{r\to\infty} \alpha_k(\omega_r)$ a.s. exists for all $k \geqq 1$. Write

$$z_k = y_{k+1} - y_k \qquad k = 1, 2, \cdots,$$

and note that by Fatou's lemma $\sum_{k=1}^{\infty} z_k \leqq 1$. This shows that $\hat{P}_n \Rightarrow \mu$ as $n \to \infty$ where μ is the distribution of the random element $z = (z_1, z_2, \cdots)$.

For $k = 1, 2, \cdots$, and $l = 1, 2, \cdots$, define the sets $A_{k,l}$ by

$$A_{k,l} = \{z_k = 0, z_{k+1} = 0, \cdots z_{k+l-1} = 0, z_{k+l} > 0\}.$$

Now for $m > r$ imagine taking a subsample of size r from the sample of m. An individual will be of the k oldest types of the smaller sample if it is one of the k oldest types in the larger sample, or if it is one of the $k-1$ oldest types of the larger sample or of the $(k+l)$th oldest type of the larger sample and none of the kth, $\cdots$, $(k+l-1)$th oldest types of the larger sample are chosen in the smaller sample. These are not the only possibilities. Thus

$$\mathbb{E}(\alpha_k(\omega_r) \mid \omega_m) \geqq \alpha_k(\omega_m) + \binom{m - m(\alpha_{k+l-1}(\omega_m) - \alpha_{k-1}(\omega_m))}{r}\binom{m}{r}^{-1}$$

$$\times \frac{\alpha_{k-1}(\omega_m) + \omega_{m,k+1}}{1 - (\alpha_{k+l-1}(\omega_m) - \alpha_{k-1}(\omega_m))}$$

where $\omega_{m,j}$ refers to the jth coordinate of the point $\omega_m \in \bar{\Delta}$. Thus writing $\beta_{k,l}(\omega_m)$ for $\alpha_{k+l-1}(\omega_m) - \alpha_{k-1}(\omega_m)$,

$$\mathbb{E}(\alpha_k(w_r)) \geqq \mathbb{E}(\alpha_k(\omega_m)) + \mathbb{E}((1 - \beta_{k,l}(\omega_m) - m^{-1}(r+1))^r$$
$$\times (1 - \beta_{k,l}(\omega_m))^{-1}(\alpha_{k-1}(\omega_m) + \omega_{m,k+l})).$$

Let $m \to \infty$ and invoke dominated convergence to obtain

$$\mathbb{E}(\alpha_k(\omega_r)) \geqq \mathbb{E}(y_k) + \mathbb{E}((1 - z_k - \cdots - z_{k+l-1})^{r-1}(y_{k-1} + z_{k+l}))$$
$$\geqq \mathbb{E}(y_k) + \mathbb{E}(1_{A_{k,l}}(1 - z_k - \cdots - z_{k+l-1})^{r-1}(y_{k-1} + z_{k+l})).$$

Now let $r \to \infty$. Dominated convergence and the definition of $A_{k,l}$ give

$$\mathbb{E}(y_k) \geqq \mathbb{E}(y_k) + \mathbb{E}(1_{A_{k,l}}(y_{k-1} + z_l)).$$

Hence $\mathbb{E}(1_{A_{k,l}} z_l) \leqq 0$, but since by definition $z_l > 0$ on $A_{k,l}$, we must have $\mathbb{P}(A_{k,l}) = 0$. However,

$$\{z \in \bar{\Delta} \backslash \Delta^*\} = \bigcup_{k=1}^{\infty} \bigcup_{l=1}^{\infty} A_{k,l}$$

so $\mathbb{P}(\{z \in \bar{\Delta} \backslash \Delta^*\}) = 0$. That is, $\mu(\Delta^*) = 1$.

Kingman (1978a), Section 4, argues convincingly that $\mu(\Delta) = 1$ in most cases of biological interest. Following him we call the measure μ which arises in the above theorem the *representing measure* of the family of consistent ordered sampling distributions and we say that families with $\mu(\Delta) = 1$ are *representable*.

https://doi.org/10.2307/1427746 Published online by Cambridge University Press

Theorem 6. *Suppose* $\{P_n\}_{n\in\mathbb{N}}$ *is a consistent family of ordered sampling distributions with representing measure* μ, *and that* $\mu(\Delta) = 1$. *Then, for all* n, *and* $\eta \in B_n$,

$$P_n(\eta) = \int_{\bar{\Delta}} \phi_\eta \, d\mu. \tag{24}$$

Furthermore this representation is unique in the sense that μ *is the only measure with* $\mu(\Delta^*) = 1$ *for which* (24) *holds.*

Proof. For $m > n$, rewrite the consistency relation (1) as $P_n(\eta) = P_n^m(\eta, \hat{P}_m)$, and now let $m \to \infty$. Theorems 1 (and this is where the requirement $\mu(\Delta) = 1$ is crucial) and (5) imply that the right-hand side converges to $\int_{\bar{\Delta}} \phi_\eta \, d\mu$ and (24) follows. The uniqueness, in the sense specified, follows from Theorem 2(b).

3.2. *Exchangeable random rankings.* For a general characterization result we need a richer environment. It is also helpful to make explicit the role played by exchangeability. Both Kingman (1982b) and Aldous (1985) have done so in the setting of partition structures. To begin, we express the problem in a slightly different form.

Say that a function $R : \mathbb{N} \times \mathbb{N} \to \{0, 1\}$ is a *ranking* if

(i) $R(i, i) = 1, \quad \forall i \in \mathbb{N}$

(ii) $R(i, j) = 0 \Rightarrow R(j, i) = 1 \quad \forall i, j \in \mathbb{N}$

(iii) $R(i, j) = 1, \quad R(j, k) = 1 \Rightarrow R(i, k) = 1 \quad \forall i, j, k \in \mathbb{N}$

$R(i, j) = 0, \quad R(j, k) = 0 \Rightarrow R(i, k) = 0 \quad \forall i, j, k \in \mathbb{N}.$

Now we imagine our infinite population consisting of individuals labelled by the positive integers. The interpretation of the function R is that $R(i, j) = 0$ if and only if individual i ranks lower than j in the ordering of the population (i.e. if $T_j > T_i$ where T_i and T_j are the types of individuals i and j respectively). Thus in the terminology of an age ordering, $R(i, j) = 0$ if i is strictly younger than j and $R(i, j) = 1$ if i is older than j or the same age as j. (Terminology is difficult here. We envisage a total order on the set of possible types. This induces an order relation on the individuals which is reflexive and transitive, but because several individuals may share the same type this induced relation is not antisymmetric and hence not a partial order.) By a random ranking we shall mean a mapping from some probability space into the set of all such rankings, with the property that each of functions $R(i, j)(\omega)$ $(i, j \in \mathbb{N})$ is measurable. A random ranking R will be called *exchangeable* if

$$(R(i, j), i, j \in \mathbb{N}) \overset{\mathcal{D}}{=} (R(\Pi(i), \Pi(j)), i, j \in \mathbb{N})$$

for any finite permutation Π on $\mathbb{N}$. (A permutation is finite if $|\{i : i \neq \Pi(i)\}| < \infty$. As usual, $|A|$ denotes the cardinality of a set A and $\overset{\mathcal{D}}{=}$ denotes equality in distribution.)

We first show that there is a one-to-one correspondence between families of consistent ordered sampling distributions and exchangeable random rankings. With

https://doi.org/10.2307/1427746 Published online by Cambridge University Press

a ranking R there is an associated equivalence relation, $\overset{R}{}$, say:

$$i \overset{R}{} j \text{ if and only if } R(i, j) = R(j, i) = 1. \tag{25}$$

Furthermore, R induces a total order on the equivalence classes of $\overset{R}{}$ in an obvious way. Now suppose R is an exchangeable random ranking. For $n = 1, 2, \cdots$, define an ordered sampling distribution $P_n(\cdot)$ by

$$P_n(l; \eta_1, \cdots, \eta_l) = P \text{ (the restriction of } \overset{R}{} \text{ to } \{1, 2, \cdots, n\} \text{ has } l \text{ equivalence classes, with the 'oldest' having } \eta_1 \text{ members, } \cdots, \text{ and the 'youngest' having } \eta_l).$$

That the family $\{P_n\}_{n \in \mathbb{N}}$ is consistent is an easy consequence of the exchangeability of R. Conversely, suppose $\{P_n\}_{n \in \mathbb{N}}$ is a consistent family of ordered sampling distributions. Let σ_n be the uniform random permutation of $\{1, 2, \cdots, n\}$ and define R^n, a random ranking of $\{1, 2, \cdots, n\}$, by insisting that with probability $P_n(l; \eta_1, \eta_2, \cdots, \eta_l)$, R^n will have l equivalence classes, the oldest of which is $\{\sigma_n(1), \cdots, \sigma_n(\eta_1)\}$, the second oldest is $\{\sigma_n(\eta_1 + 1), \cdots, \sigma_n(\eta_1 + \eta_2)\}, \cdots$, and the youngest is $\{\sigma_n(\eta_1 + \cdots + \eta_{l-1}), \cdots, \sigma_n(n)\}$. Consistency of $\{P_n\}$ and properties of σ_n ensure that for each n, R^n is a (finitely) exchangeable ranking of $\{1, 2, \cdots, n\}$ and that

$$(R^{n+1}(i, j) : i, j \in \{1, 2, \cdots, n\}) \overset{\mathcal{D}}{=} (R^n(i, j) : i, j \in \{1, 2, \cdots, n\}).$$

The Daniel–Kolmogorov theorem then guarantees the existence of an exchangeable random ranking R (of $\mathbb{N}$) with the property that

$$(R(i, j) : i, j \in \{1, \cdots, n\}) \overset{\mathcal{D}}{=} (R^n(i, j) : i, j \in \{1, 2, \cdots, n\}),$$

for each $n \geqq 1$.

A characterization of exchangeable random rankings is thus equivalent to a characterization of consistent ordered sampling distributions. We give the former, proving an analogue of de Finetti's theorem in this setting. The role of independent and identically distributed sequences is played by the following 'paintbox process'. Let μ be a probability measure on $[0, 1]$, and let $\xi_1, \xi_2, \cdots$ be i.i.d. μ. Define a random ranking R by

$$R(i, j) = 1 \text{ if and only if } \xi_i \geqq \xi_j. \tag{26}$$

Clearly R is exchangeable. We call it the paintbox process with distribution μ. The terminology (Kingman (1982b), Aldous (1985)) stems from the idea of using the real numbers $0 \leqq x \leqq 1$ to label the colours of the spectrum, and imagining a collection of objects $1, 2, \cdots$ being coloured at random by painting object i with colour ξ_i. The equivalence classes of R consist of those objects which are identically coloured. (Observe that with probability 1 these are either infinite, in which case they correspond to atoms of μ, or singletons.) The equivalence classes here are ordered by their colours, or equivalently by the usual order on $[0, 1]$. Note that the

https://doi.org/10.2307/1427746 Published online by Cambridge University Press

distribution of R depends only on the sizes of the ordered atoms of μ and on the ordered collection of masses between neighbouring atoms. We denote the distribution of R by P^{μ}.

It should now be clear that the characterization results of Section 3.1 cannot be extended beyond the representable case because different paintbox processes (and hence different families of consistent ordered sampling distributions) share the same representing measure.

Before proving the general result, we collect together some of the properties enjoyed by exchangeable random rankings. We abuse terminology and refer to the equivalence classes of the induced equivalence relation (25) as (equivalence) classes of the ranking itself.

Proposition 7. *Let R be an exchangeable random ranking with classes*

$$C_i \equiv \{k : R(i, k) = R(k, i) = 1\}, \qquad i = 1, 2, \cdots$$

(i) *For each i, $P(|C_i| = 1$ or $|C_i| = \infty) = 1$. On the set $\{|C_i| = \infty\}$,*

$$\lim_{n \to \infty} n^{-1} |C_i \cap \{1, 2, \cdots, n\}| = W_1 \tag{27}$$

exists almost surely and $P(W_1 > 0 \mid |C_i| = \infty) = 1$.

(ii) *Let $S \equiv \{i : |C_i| = 1\}$ denote the set of singleton classes in R. Then $P(|S| = 0$ or $|S| = \infty) = 1$ and on the set $\{|S| = \infty\}$,*

$$\lim_{n \to \infty} n^{-1} |S \cap \{1, 2, \cdots, n\}| = W_2$$

exists almost surely, and $P(W_2 > 0 \mid |S| = \infty) = 1$.

(iii) *Let $j \in \mathbb{N}$ be fixed and denote by S_j^+ and S_j^- the set of singleton classes in R which are ranked above and below j respectively*:

$$S_j^+ = \{i \in S : R(j, i) = 0\} \qquad S_j^- = \{i \in S : R(i, j) = 0\}.$$

Then $P(|S_j^+| = 0$ or $|S_j^+| = \infty) = 1$ and on the set $\{|S_j^+| = \infty\}$

$$\lim_{n \to \infty} n^{-1} |S_j^+ \cap \{1, 2, \cdots, n\}| = W_3^+$$

exists almost surely, and $P(W_3^+ > 0 \mid |S_j^+| = \infty) = 1$. Analogous statements hold for S_j^-.

(iv) *Let $j \in \mathbb{N}$ be fixed. Suppose $|S| > 0$ and let $i \in S$, $i \neq j$, correspond to a singleton class. Let A_{ij} denote the set of singleton classes ranked between i and j*:

$$A_{ij} = \{k \in S : R(i, k) = 0 \text{ and } R(k, j) = 0\} \cup \{k \in S : R(j, k) = 0 \text{ and } R(k, i) = 0\}.$$

Then

$$\lim_{n \to \infty} n^{-1} |A_{ij} \cup \{1, 2, \cdots, n\}| = W_4$$

exists almost surely, and $P(W_4 > 0 \mid |S| > 0) = 1$.

https://doi.org/10.2307/1427746 Published online by Cambridge University Press

Proof. First note that if R is an exchangeable random ranking, then $\overset{R}{\sim}$ the equivalence relation induced by R is an exchangeable random equivalence relation. Statements (i) and (ii) concern the equivalence classes of $\overset{R}{\sim}$ and the results follow from Kingman's (1982b) characterization of exchangeable random equivalence relations. For (iii), define a sequence $X_1, X_2, \cdots, X_{j-1}, X_{j+1}, \cdots$ of $\{0,1\}$-valued random variables by

$$X_i = \begin{cases} 1, & \text{if } i \in S \text{ and } R(j,i) = 0 \\ 0, & \text{otherwise.} \end{cases}$$

It is straightforward to check that this sequence inherits exchangeability from R, and the result for S_j^+ is then a consequence of de Finetti's theorem. The result for S_j^- follows *mutatis mutandis*.

We prove (iv) in the case $R(j,i) = 0$. The other case is similar. Let $a_n = |S_j^+ \cap \{1, 2, \cdots, n\}|$ and $b_n = |A_{ij} \cap \{1, 2, \cdots, n\}|$. By exchangeability, the distribution of b_n conditional on a_n is uniform on $\{0, 1, \cdots, a_n - 1\}$. Since $|S_j^+| \geqq 1$, it follows from (iii) that $a_n \to \infty$ a.s. and hence that

$$b_n/a_n \Rightarrow V, \tag{28}$$

say, where V has a uniform distribution on $[0,1]$. Again by exchangeability

$$E(b_n/a_n \mid a_m, b_m, a_{m+1}, b_{m+1}, \cdots) = b_m/a_m,$$

so $\{b_n/a_n; n = 1, 2, \cdots\}$ is a reversed martingale (bounded above by 1) from which it follows that $b_n/a_n \to V$ almost surely for some V. It follows from (28) that V has a uniform distribution on $[0,1]$, and hence in particular that $P(V > 0) = 1$.

Now by assumption $|S_j^+| \geqq 1$ so from (iii), $|S_j^+| = \infty$ a.s. and

$$\begin{aligned} \lim_{n\to\infty} n^{-1} |A_{ij} \cap \{1, 2, \cdots, n\}| &= \lim_{n\to\infty} \frac{b_n}{a_n} \frac{a_n}{n} \\ &= VW_3^+ \end{aligned}$$

almost surely. Writing $W_4 = VW_3^+$ we have

$$P(W_4 = 0) \leqq P(V = 0) + P(W_3^+ = 0) = 0$$

which completes the proof of (iv).

We can now characterize any exchangeable random ranking as a mixture of paintbox processes.

Theorem 8. *Let R be an exchangeable random ranking. There exists a random probability measure $\tilde{\mu}$ on $[0,1]$ for which the conditional distribution of R given $\tilde{\mu} = \mu$ is P^μ, the paintbox process with distribution μ.*

Proof. The sequence $\{R(i,j) : j = 1, 2, \cdots, j \neq i\}$ is an exchangeable sequence of

https://doi.org/10.2307/1427746 Published online by Cambridge University Press

zeros and ones. Define p_i to be the limiting proportion of ones:

$$p_i = \lim_{n\to\infty} n^{-1} |\{j \leqq n, j \neq i : R(i, j) = 1\}|$$

which exists (almost surely) by de Finetti's theorem. Observe that we can write $(p_1, p_2, \cdots) = f(R(i, j), i, j \in \mathbb{N})$ for a particular function f, and furthermore that if Π is a finite permutation on $\mathbb{N}$, then $f(R(i, j)) \overset{\mathcal{D}}{=} f(R(i, \Pi(j)))$.

Now for such a permutation

$$(p_{\Pi(1)}, p_{\Pi(2)}, \cdots) = f(R(\Pi(i), (j)) = f(R(\Pi(i), \Pi(j)))$$
$$\overset{\mathcal{D}}{=} f(R(i, j)) = (p_1, p_2, \cdots),$$

so the sequence $p_1, p_2, \cdots$ is exchangeable. Invoke de Finetti's theorem once more and denote by $\tilde{\mu}$ the random directing measure of this sequence: conditional on $\tilde{\mu} = \mu$, $p_1, p_2, \cdots$ are i.i.d. μ. We claim that conditional on $\tilde{\mu} = \mu$, R has the distribution of a paintbox process with distribution μ. In fact we show that $R(i, j) = 1$ if and only if $p_i \geqq p_j$.

First suppose $R(i, j) = 1$. By transitivity

$$\{k \leqq n : R(j, k) = 1\} \subseteq \{k \leqq n : R(i, k) = 1\}$$

so clearly $p_i \geqq p_j$. A similar argument shows that $R(i, j) = 0 \Rightarrow p_i \leqq p_j$, so $p_i > p_j \Rightarrow R(i, j) = 1$. It only remains to show that $p_i = p_j \Rightarrow R(i, j) = 1$.

Suppose $p_i = p_j$ but $R(j, k) = 0$. Let $B_{ij} = \{k : R(i, k) = 0, R(j, k) = 1\}$. Then

$$0 = p_j - p_i = \lim_{n\to\infty} n^{-1} |B_{ij} \cap \{1, 2, \cdots, n\}|. \tag{29}$$

As in Proposition 7, denote by C_i and C_j respectively the classes containing i and j, and note that $R(i, j) = 0 \Rightarrow C_i \neq C_j$. From Proposition 7(i), with probability 1 either $|C_i| = |C_j| = \infty$ or at least one of C_i or C_j is a singleton. In the first case $C_j \subseteq B_{ij}$ and (27) and $P(W_1 > 0)$ gives a contradiction to (29). In the second case apply Proposition 7(iv) and observe that $A_{ij} \subseteq B_{ij}$ to obtain a contradiction.

Remark 1. This technique provides another proof of Kingman's representation of partition structures (1978b) or equivalently of exchangeable equivalence relations (1982b). The original proof, using martingale techniques, was similar in spirit to the arguments of Section 3.1, and to some proofs of de Finetti's theorem (see for example Aldous (1985), pp. 21, 22). Aldous (1985) provides an elegant proof using the device of 'external randomization' and another by exploiting deep results on partial exchangeability.

Remark 2. In the unordered case the extension from the representable case simply has to cope with the possibility of singleton classes. Allowing representing measures on $\bar{\Delta}$, with $1 - \mu(\Delta)$ representing the proportion of singleton classes, is sufficient for this.

https://doi.org/10.2307/1427746 Published online by Cambridge University Press

Remark 3. The class, $\mathscr{C}$ say, of all exchangeable random rankings, or more precisely of their distributions, or equivalently the class of families of consistent ordered sampling distributions, may be regarded as a particular projective limit, with the projections arising from (1) for ordered sampling distributions and restriction (to $\{1, 2, \cdots, n\}$) for exchangeable random rankings (of $\{1, 2, \cdots, m\}$, $m > n$). As a projective limit, the class $\mathscr{C}$ is (Davies and Vincent-Smith (1968)) a Choquet simplex. Within this framework the characterization problem addressed here may be translated as the problem of describing the simplex $\mathscr{C}$, or as is by Choquet theory the same, of identifying its extreme points. We have proved that the extreme points may be labelled by the equivalence classes of the set of probability measures on $[0, 1]$ (with two such measures being equivalent if they give rise to paintbox processes with the same distribution), and that these extreme points are exactly the (distributions of) paintbox processes. In a related setting this is (like partition structures) a concrete example of the theory of Lauritzen (1974) in whose terminology the paintbox processes would be called 'extreme models'.

4. Convergence of age-ordered population frequencies in genetics

As has already been mentioned, ordered sampling distributions arise naturally in population genetics. In haploid, single-locus, models we can associate a unique genetic type, or allele, with each individual in the population. New alleles arise by mutation and at any time the collection of alleles present in the population may be ordered by age, that is time since their first appearance in the population. Furthermore for many models, at least at equilibrium and in the so-called diffusion limit (that is as the population size becomes large) the ordering by age is well defined. (Ties in ages occur with probability zero.) Sampling distributions are of genuine interest because it is in this form that the geneticist actually obtains data.

If information about ages is either unavailable or of no interest then under certain conditions the Ewens sampling formula (Ewens (1972)) provides a robust description of samples from a wide range of populations at equilibrium. Kingman (1977) has related this sampling distribution to a Poisson–Dirichlet population structure. For problems where the age structure is of interest an analogue of the Ewens sampling formula has recently been found (Donnelly and Tavaré (1986)). Our purpose here is exhibit the population structure implied by this family of ordered sampling distributions and to exploit the convergence of sampling distributions (as the population size $N \to \infty$) for a range of discrete models, to prove (in the spirit of Section 2) convergence of age-ordered population frequencies.

We first describe a general model, initially introduced by Cannings (1974) for the reproduction of a haploid population. Assume a population evolves in discrete, non-overlapping generations, and that each generation is of fixed size N. (In fact as noted in Kingman (1982a) and elsewhere, the following results apply to a number of models with overlapping generations, notably the Moran model, and varying

https://doi.org/10.2307/1427746 Published online by Cambridge University Press

population sizes.) Label the members of the rth generation in random order, and let ν_i be the number of offspring born to the ith member. The random variables $\nu_1, \nu_2, \cdots, \nu_N$ are then exchangeable and $\nu_1 + \nu_2 + \cdots + \nu_N = N$. Further assume that the ν_j are independent of family sizes in other generations (this is an assumption of genetic neutrality) and that the joint distribution of the ν_j does not depend on r. The most celebrated model of population genetics, the Wright–Fisher model, corresponds to the case in which the joint distribution of these offspring numbers is symmetric multinomial. Introduce mutation by assuming that an offspring adopts the type of its parent with probability $(1-u)$ and otherwise is of a novel type, never before seen in the population, all such choices being independent of each other and of the past history of the population. (This is the so-called infinite alleles assumption.) Throughout the following we shall also assume that the reproductive model satisfies the following moment conditions:

$$\begin{aligned} &\text{(i)}\ \operatorname{var}(\nu_1) \to \sigma^2 > 0 \text{ as } N \to \infty \\ (30)\quad &\text{(ii)}\ \sup_N E(\nu_1^k) < \infty, \quad k = 1, 2, \cdots \\ &\text{(iii)}\ u = \frac{\theta\sigma^2}{2N} + o(N^{-1}), \text{ for some } \theta > 0 \text{ as } N \to \infty. \end{aligned}$$

Focus attention on a particular exchangeable model and a given population size N. After t generations there will be a certain number of alleles present. Label these in decreasing order of age, so that allele 1 is the oldest in the population, allele 2 the second oldest, and so on. For convenience adopt the convention that if two or more alleles share a particular age (because they arose as novel mutations in the same generation) they will be ordered at random. Denote by $Y_N(t)$ $(\in \Delta_N)$ the proportion of the population of each of these types, ordered by age. It follows from the form of the reproductive model that $\{Y_N(t); t = 0, 1, \cdots\}$ is an irreducible Markov chain with state space Δ_N, and since this is finite, there exists a unique equilibrium distribution which we denote by λ_N.

Considerable progress in the analysis of such models has recently been made by concentrating on the genealogy which they induce. In particular if one studies the genealogy of a sample of fixed size, n, say, then with an appropriate time scaling (by $N\sigma^{-2}$ in the notation of (30)) the *genealogical process* associated with any of the exchangeable models satisfying (30) converges weakly as $N \to \infty$ to a particularly tractable limiting process called an n-coalescent with ages. (See Kingman (1982c) in the case of no mutation. The extension to allow for mutation and the age ordering is due to Donnelly and Tavaré (1986).) The fact that the sample size is fixed as the population size increases is crucial to the proofs of these results, so that they do not immediately give any information about population behaviour. (Donnelly (1991) gives an extension to *population* ancestral processes, but that is in some sense complementary to the result of interest here.) One consequence of the main result

https://doi.org/10.2307/1427746 Published online by Cambridge University Press

of Donnelly and Tavaré (1986) is convergence of age-ordered sampling distributions at equilibrium. For convenience we restate that part of the result which is needed here.

Theorem 9. *Denote by* λ_N *the equilibrium distribution of age-ordered population frequencies in one of the exchangeable models satisfying* (30), *and as in* (2) *by* $P_n^N(\cdot, \lambda_N)$ *the equilibrium distribution of age-ordered samples of size n. Then as* $N \to \infty$, *for each* $\eta \in B_n$,

$$P_n^N(\eta, \lambda_N) \to P_n(\eta) \tag{31}$$

where (*here and subsequently*) $P_n(\cdot)$ *is given by*

$$P_n((l; \eta_1, \cdots, \eta_l)) = \frac{\theta^l}{\theta_{(n)}} \frac{n!}{\eta_l(\eta_l + \eta_{l-1}) \cdots (\eta_l + \eta_{l-1} + \cdots + \eta_1)} \tag{32}$$

where $\theta_{(n)} = \theta(\theta + 1) \cdots (\theta + n - 1)$.

It follows from Theorem 3 that $\{P_n\}_{n \in \mathbb{N}}$ is a consistent family of ordered sampling distributions. (Verification of consistency is in any case merely an exercise in algebra.) It is natural to ask about its representing measure.

Let $Z_1, Z_2, \cdots$ be a collection of independent and identically distributed random variables with common probability density function $f(z) = \theta(1 - z)^{\theta - 1}$, $0 \leqq z \leqq 1$. Construct a random element X of $\bar{\Delta}$ by putting $X = (Z_1, (1 - Z_1)Z_2, (1 - Z_1)(1 - Z_2)Z_3, \cdots)$, and define a measure λ on $\bar{\Delta}$ by putting $\lambda(A) = P(X \in A)$ for Borel sets A. Note that $\lambda(\Delta) = 1$. The measure λ is the so called Griffiths–Engen–McCloskey, or GEM distribution. Its properties are quite remarkable (see for example Donnelly (1986) or Ewens (1990) for an exploitation of these) and fully justify the acronym.

Theorem 10. *The measure* λ *corresponding to the GEM distribution is the representing measure of the family* $\{P_n\}_{n \in \mathbb{N}}$ *and*

$$P_n(\eta) = \int_{\bar{\Delta}} \phi_\eta \, d\lambda. \tag{33}$$

Proof. Donnelly and Tavaré ((1986), Proposition 6.1) prove that if $\eta^{(n)}$ has the distribution P_n given by (32) then for fixed r, the joint distribution of $(n^{-1}\eta_1^{(n)}, n^{-1}\eta_2^{(n)}, \cdots, n^{-1}\eta_r^{(n)})$ converges to that of $(Z_1, (1 - Z_1)Z_2, \cdots, (1 - Z_1)(1 - Z_2) \cdots (1 - Z_{r-1})Z_r)$. It follows from this and the fact that the functions which depend on only a finite number of coordinates are dense in $C(\bar{\Delta})$, the continuous functions from $\bar{\Delta}$ to $\mathbb{R}$, that (in the notation of (9)) $\hat{P}_n \Rightarrow \lambda$. From Theorem 5 we know that $\hat{P}_n \Rightarrow \mu$, the representing measure of $\{P_n\}_{n \in \mathbb{N}}$, so $\mu = \lambda$. Since $\lambda(\Delta) = 1$, (33) follows from Theorem 6.

Theorem 11. *Denote by* λ_N *the equilibrium distribution of age-ordered population frequencies in a population of size N evolving according to a particular exchangeable model satisfying* (30). *Then if* ν *denotes a limit point of the sequence* $\{\lambda_N\}$, $\nu(\{x : x_i > 0 \ \forall i\}) = 1$.

https://doi.org/10.2307/1427746 Published online by Cambridge University Press

The proof of the theorem requires a more detailed analysis of the time evolution of genealogical processes associated both with samples and with the whole population, and we defer it to the end of the section.

Theorem 12. *Denote by* λ_N *the equilibrium distribution of age-ordered population frequencies for a population of size N evolving according to a particular exchangeable model satisfying* (30). *Then as* $N \to \infty$, $\lambda_N \Rightarrow \lambda$, *where* λ *is the measure associated with the GEM distribution.*

Proof. In the light of (31), (33), and Theorem 11, the result is an immediate consequence of Theorem 4.

Remark 1. The key ingredients used to prove Theorem 12 were the convergence of ordered sampling distributions (31) and the fact that limit points of the sequence $\{\lambda_N\}$ place mass 1 on $\{x : x_i > 0 \, \forall i\}$. As mentioned at the beginning of the section, the convergence (31) holds for a class of models somewhat wider than those described in Theorem 9 so that if Theorem 11 could be proved for one such model, then Theorem 12 would also be valid for that model. The proof of Theorem 11 rests heavily on convergence of non-equilibrium sampling distributions. This too is likely to be true for other models (those for which (31) holds), in which case the required result should follow by the same means.

Remark 2. For one specific model not encompassed by the conditions of Theorem 12, the Moran model, the equilibrium distribution λ_N is known explicitly (Donnelly and Tavaré (1986)). It has the form $\hat{P}_N$ (with P_N given by (32) with $n = N$.) The result $\lambda_N \Rightarrow \lambda$ then follows immediately from the fact that $\hat{P}_n \Rightarrow \lambda$.

Remark 3. One could imagine proving Theorem 12 directly for a particular model, say the Wright–Fisher model. Such a scheme would involve explicit calculation of λ_N, or a good approximation of it, or possibily of the whole (forward) process for which λ_N is the equilibrium distribution, and in the spirit of, say, Ethier and Kurtz (1986), Chapter 10, is likely to be highly model-specific. One advantage of the genealogical approach, here and elsewhere, is that the results apply to a wide range of models. At the sample level (Kingman (1982c), Donnelly and Tavaré (1986)) this involves few model-specific calculations, and the transition here to population results involves *no* further model-specific calculations.

Remark 4. Kingman (1977) effectively proves the analogue of Theorem 12 in the case where age ordering is of no interest. The population distribution there is the Poisson–Dirichlet and the route is rather easier because populations may be described by points of the subset ∇ of Δ consisting of sequences whose components are non-increasing.

Proof of Theorem 11. Throughout, we consider a fixed exchangeable reproductive model satisfying (30). The idea is to study the population genealogy associated with the model, and to exploit known results to prove convergence of the marginal

https://doi.org/10.2307/1427746 Published online by Cambridge University Press

distribution of an appropriate genealogical process. For a given allele in the population at equilibrium this gives information about the behaviour of the population since that allele first arose, and hence information about its frequency.

We first recall the appropriate definitions and facts. For a more leisurely account see Donnelly and Tavaré (1986) for sample results and Donnelly and Tavaré (1987) for population genealogy. Fix some reference time (and call this time 0) and consider the composition of the population with respect to the ancestral population t generations into the past. Say that a pair of individuals are in the same old class if they share a common ancestor in generation $-t$ and there is no mutation in their lines of descent from the common ancestor to the present. Say that they are in the same new class if they share a common ancestor in generation $-r$, $0 < r \leqq t$, where that ancestor itself is a mutant and there is no further mutation in their lines of descent from this common ancestor to the present. This partitions the population of N (or a sample from it) into a collection of old and new classes. All the individuals in a given (old or new) class will be of the same allelic type and (because of the infinite alleles assumption) different new classes correspond to different types. The new classes may then be ordered by age (of the corresponding allele), and again we suppose that ties are broken at random. If there are K old classes with sizes $a_1 > a_2 > \cdots > a_K$ say, and F new classes with sizes $b_1, b_2, \cdots, b_F$, say, written in age order (so that the type of the new class of size b_1 is the oldest, that of the b_2 the second oldest and so on) the genealogy of the population may be described by a point

$$M^N(t) = (N^{-1}a_1, N^{-1}a_2, \cdots, N^{-1}a_K; N^{-1}b_1, N^{-1}b_2, \cdots, N^{-1}b_F). \tag{34}$$

As t increases this gives rise to a stochastic (in fact Markov) process. Note that changes in the process take one of two forms: either two or more of the old classes coalesce (due to a sharing of ancestors) or one of the old classes becomes the oldest of the new classes (because the ancestor of the old class is a mutant). In the limit as $t \to \infty$, all of the classes will be new, and furthermore the distribution of age-ordered class frequencies will be given by λ_N, the equilibrium distribution of the forward process of allele frequencies with age ordering. (This last in particular and the whole genealogical approach in general corresponds to a form of duality.) In order to show that limit points of λ_N concentrate on $\{x : x_i > 0 \ \forall i\}$, for each l we effectively bound $N^{-1}b_l$ below in probability (uniformly in N) by bounding the old class sizes prior to the time when that particular new class was formed.

We now recall some known results. Denote by $A^N(t)$ the number of old classes in $M^N([N\sigma^{-2}t])$ (where $[y]$ denotes the integer part of y) and by $D_\infty(\cdot)$ the Markov death processes on $E = \{0, 1, 2, \cdots, \infty\}$ with entrance boundary (and initial value) at ∞ and death rates $k(k + \theta - 1)/2$ from state k. Then (Donnelly (1991), Section 5, where $A^N(\cdot)$ is called $B_N^N(\cdot)$) as $N \to \infty$, $A^N(\cdot) \Rightarrow D_\infty(\cdot)$ as elements of $D_E[0, \infty)$. Further suppose a sample of size n (in which the individuals are labelled from the set $\{1, 2, \cdots, n\}$) is taken from the population at time 0, and its genealogy with respect

https://doi.org/10.2307/1427746 Published online by Cambridge University Press

to the ancestral population at time t is described by listing the individuals in the old and (age-ordered) new classes. That is, $R_n^N(t) = (\xi_1, \cdots, \xi_k; \zeta_1, \cdots, \zeta_l)$ if there are k old classes (written in arbitrary order) the ith of which consists of the individuals with labels in the subset $\xi_i \subseteq \{1, 2, \cdots, n\}$ $i = 1, 2, \cdots, k$ and l new classes with the individuals in ζ_1 having the oldest new type, those in ζ_2 the second oldest, and so on. (Denote by $\mathscr{E}_n$ the state space of $R_n^N(\cdot)$ and observe that it is finite.) The central result of Donnelly and Tavaré (1986) ensures that if the population is evolving according to one of the exchangeable models satisfying the conditions (30), then as $N \to \infty$

$$R_n^N([N\sigma^{-2}\cdot]) \Rightarrow R_n(\cdot) \tag{35}$$

(as elements of $D_{\mathscr{E}_n}[0, \infty)$) where $\{R_n(t); t \geqq 0\}$ is a particularly tractable process called an n-coalescent with ages.

Now focus attention on the old classes of the process $M^N(\cdot)$ defined at (34). Specifically, define a process $X^N(t)$ by $X^N(t) = \tau(M^N([N\sigma^{-2}t]))$, where τ is defined by

$$\tau(x_1, x_2, \cdots, x_k; y_1, y_2, \cdots, y_F) = (x_1, x_2, \cdots, x_k, 0, 0, \cdots).$$

Recall from (34) that $x_1 \geqq x_2 \geqq \cdots \geqq x_k$.

The appropriate limiting process $X(\cdot)$ is the restriction to old classes (rearranged in descending order of class sizes) of the infinite population coalescent with ages of Donnelly and Tavaré (1987). For our purposes it suffices to observe that $X(\cdot)$ takes values in the space $\bar{\nabla} = \{x \in \bar{\Delta} : x_1 \geqq x_2 \geqq x_3 \geqq \cdots\}$, (which we topologize as a closed subset of $\bar{\Delta}$), and that the distribution of $X(t)$ has the following representation. For $k = 1, 2, \cdots$, let V_k be a random variable with probability density function

$$f_k(v) = \frac{\Gamma(k+\theta)}{\Gamma(k)\Gamma(\theta)} v^{k-1}(1-v)^{\theta-1}, \qquad 0 \leqq v \leqq 1.$$

Independent of V_k, let $U^{(k)}$ have a uniform distribution on the simplex $\{(x_1, \cdots, x_k) : x_i \geqq 0, x_1 + \cdots + x_k = 1\}$ and let $U_{(1)}^{(k)}, U_{(2)}^{(k)}, \cdots, U_{(k)}^{(k)}$ denote the descending order statistics of its components. Then $X(0) = (0, 0, \cdots)$ and for a Borel subset A of $\bar{\Delta}$,

$$P(X(t) \in A) = P(D_\infty(t) = 0) 1_{\{(0,0,\cdots) \in A\}} + \sum_{k=1}^{\infty} P(D_\infty(t) = k) 1_{\{V_k U_{(1)}^{(k)}, V_k U_{(2)}^{(k)}, \cdots, V_k U_{(k)}^{(k)}, 0, \cdots) \in A\}}$$

for $t > 0$.

Proposition 13. *For each $t > 0$, as $N \to \infty$, $X^N(t) \Rightarrow X(t)$ as random elements of $\bar{\nabla}$.*

Proof. Denote by $\mathscr{B}_n$ the set of subpartitions of the integer n: $\mathscr{B}_n = \{A = \{a_1, \cdots, a_l\} : a_i \in \mathbb{N}, a_1 + \cdots + a_l \leqq n\}$ and by $\tilde{\phi}_A^n : \bar{\nabla} \to [0, 1]$ the sampling function

$$\tilde{\phi}_A^n(x) = \binom{n}{a_0} \left(1 - \sum_{k=1}^{\infty} x_k\right)^{a_0} \frac{(n-a_0)!}{a_1! \cdots a_l!\, \beta_1! \cdots \beta_n!} \sum x_{i_1}^{a_1} x_{i_2}^{a_2} \cdots x_{i_l}^{a_l}$$

https://doi.org/10.2307/1427746 Published online by Cambridge University Press

where $A = \{a_1, \cdots, a_l\}$, $a_0 = n - a_1 - \cdots - a_l$ and for $j = 1, 2, \cdots, n$, β_j is equal to the number of a_i $(i = 1, \cdots, l)$ which take the value j. The second summation is over all l-tuples of distinct positive integers $i_1, \cdots, i_l$. Endow $\mathcal{B}_n$ with the discrete topology.

Our present concern is only with the sizes of the old classes of the sample coalescent R_n^N, so define the restriction map $\tilde{\tau}: \mathcal{E}_n \to \mathcal{B}_n$ by

$$\tilde{\tau}(\xi_1, \cdots, \xi_k; \zeta_1, \zeta_2, \cdots, \zeta_l) = \{|\xi_1|, |\xi_2|, \cdots, |\xi_k|\}$$

where $|\xi_i|$ denotes the number of elements in the class ξ_i. Write $W_n^N(t)$ for $\tilde{\tau}R_n^N([N\sigma^{-2}t])$ and $W_n(t)$ for $\tilde{\tau}R_n(t)$. The continuous mapping theorem and (35) guarantee that as $N \to \infty$, $W_n^N(\cdot) \Rightarrow W_n(\cdot)$ as elements of $D_{\mathcal{B}_n}[0, \infty)$. Observe that the set of fixed jumps of W_n is empty, that is

$$T_W = \{0\} \cup \{t > 0 : P(W_n(t-) = W_n(t)) = 1\} = [0, \infty).$$

The fact that W_n^N describes the genealogy of a sample of size n from the population of N, and calculations analogous to those in the proof of Theorem 1 give, for any $A \in \mathcal{B}_n$,

$$C_N E[\tilde{\phi}_A^n(X^N(t))] - H_N \leqq P(W_n^N(t) = A) \leqq C_N E[\tilde{\phi}_A^n(X^N(t))] \tag{36}$$

for constants C_N and H_N (depending on N) with the property that $C_N \to 1$, $H_N \to 0$ as $N \to \infty$. Thus, from (36)

$$\lim_{N \to \infty} E[\tilde{\phi}_A^n(X^N(t))] = P(W_n(t) = A).$$

The distribution of $W_n(t)$ is given in Watterson (1984), Equation 3.4.1. A straightforward calculation, which we omit in the interests of brevity, shows that

$$P(W_n(t) = A) = E[\tilde{\phi}_A^n(X(t))].$$

(The key to performing the calculation of the right-hand side is first to condition on the number of old classes in $X(t)$, proceeding as in the derivation of Equation (3.19) of Kingman (1982b), and then to relate the distribution of the number of old classes in $X(t)$ to the number of old classes in $R_n(t)$, via Equation (7.15) of Tavaré (1984).) Hence

$$\lim_{n \to \infty} E[\tilde{\phi}_A^n(X^N(t))] = E[\tilde{\phi}_A^n(X(t))]. \tag{37}$$

Kingman (1977) shows that a subset of the family $\{\phi_A^n, A \in \mathcal{B}_n, n = 1, 2, \cdots\}$ generates an algebra which is dense in $C(\bar{\nabla})$, the continuous functions from $\bar{\nabla}$ into R, so (37) ensures $X_N(t) \Rightarrow X(t)$ as required.

Now suppose ν is a limit point of the sequence $\{\lambda_N\}$. For convenience, index the relevant convergent subsequence by N. Thus in what follows $\lambda_N \Rightarrow \nu$ by assumption.

https://doi.org/10.2307/1427746 Published online by Cambridge University Press

We show that for any i, $\nu(\{x_i > 0\}) = 1$, and hence that

$$\nu(\{x_i > 0 \ \forall i\}) = 1 - \nu\left(\bigcup_i \{x : x_i = 0\}\right) \geqq 1 - \sum_{i=1}^{\infty} \nu(\{x : x_i = 0\}) = 1.$$

First observe that

$$\nu(\{x_i > 0\}) = 1 \text{ if only if for each } \varepsilon > 0, \ \exists\, \delta > 0 \text{ with } \nu(\{x_i \geqq \delta\}) > 1 - \varepsilon.$$

Now the set $\{x : x_i \geqq \delta\}$ is closed in $\bar{\Delta}$, so by the portmanteau theorem it is enough to establish that given $\varepsilon > 0$, $\exists\, \delta > 0$ with

$$\limsup_{N \to \infty} \lambda_N(\{x_i \geqq \delta\}) > 1 - \varepsilon.$$

Fix a value of t, to be specified presently. The central idea in what follows is that if t is small enough, then with high probability the ith new class (at equilibrium) will have been an old class at time t in the genealogical process $X^N(\cdot)$, and so its size is bounded below by the size of the smallest of the old classes at time t, which in turn may be bounded away from zero with high probability.

Define the random variables

Q_k^N = the number of new classes formed in a k-coalescent with ages from the population of size N,

$C^N(t)$ = the number of new classes formed from time $[N\sigma^{-2}t]$ onwards in $M^N(\cdot)$.

Conditional on $A^N(t) = k$, the k ancestors of the old classes in $M^N([N\sigma^{-2}t])$ are equally likely to be any of the $\binom{N}{k}$ possible subsets of ancestors, and the number of new classes formed from $[N\sigma^{-2}t]$ onwards is a function only of the genealogy of these k ancestors. Thus, conditional on $A^N(t) = k$, the distribution of $C^N(t)$ is that of Q_k^N, which does not depend on t. Note further that the number of new classes formed in $M^N(\cdot)$ from time $[N\sigma^{-2}t]$ onwards is a function of the number of old classes in $M^N([N\sigma^{-2}t])$ and not on their sizes, so that $C^N(t)$ and $X^N(t)$ are conditionally independent given $A^N(t)$.

Write $X_i^N(t)$ $i = 1, 2, \cdots$ for the components of $X^N(t)$ and recall that these components are written in decreasing order. Define the function $\Lambda : \bar{\nabla} \to [0, 1]$ by $\Lambda((x_1, x_2, \cdots)) = \inf\{x_i : x_i > 0\}$, (with $\inf \phi = 0$) so that in particular if x has only a finite number of non-zero components, then $\Lambda(x)$ is the smallest of these. Then in view of the comments in the preceding paragraph and the fact that on $\{A^N(t) = k\}$, $\Lambda(X^N(t)) = X_k^N(t)$,

$$\begin{aligned}
\lambda_N(\{x_i \geqq \delta\}) &\geqq P(\Lambda(X^N(t)) \geqq \delta,\ C^N(t) \geqq i) \\
&= \sum_{k=0}^{N} P(A^N(t) = k) P(C^N(t) \geqq i \mid A^N(t) = k) \\
&\quad \times P(\Lambda(X^N(t)) \geqq \delta \mid A^N(t) = k,\ C^N(t) \geqq i) \\
&= \sum_{k=0}^{N} P(A^N(t) = k) P(Q_k^N \geqq i) P(X_k^N(t) \geqq \delta \mid A^N(t) = k).
\end{aligned} \tag{38}$$

https://doi.org/10.2307/1427746 Published online by Cambridge University Press

Now suppose $k > l$. The genealogical process $\tilde{R}_l(\cdot)$, say, associated with a sample of size l may be naturally embedded in (i.e. coupled with) that associated with a sample of size k, $\tilde{R}_k(\cdot)$, by restricting attention to the first l elements of the larger sample. (This is Kingman's (1982a) natural coupling.) In this coupling the number of new classes formed in $\tilde{R}_l(\cdot)$ is not larger than the number of new classes formed in $\tilde{R}_k(\cdot)$. Thus the random variables Q_k^N are stochastically non-decreasing in k. It follows from (38) that for any $N > L > K \geqq 0$,

$$\begin{aligned}\lambda_N(\{x_i \geqq \delta\}) &\geqq P(Q_K^N \geqq i) \sum_{k=K}^{L} P(A^N(t) = k) P(X_k^N(t) \geqq \delta \mid A^N(t) = k) \\ &= P(Q_K^N \geqq i) \sum_{k=K}^{L} P(X_k^N(t) \geqq \delta,\ X_{k+1}^N(t) = 0).\end{aligned} \tag{39}$$

Now, since the sets $\{x_k \geqq \delta, x_{k-1} > 0\}$ each have boundary contained in the set $\{x_k = \delta\}$, an event which (provided $\delta > 0$) has probability zero under the distribution of $X(t)$, we may invoke Proposition 13 to write

$$\begin{aligned}P(X_k^N(t) \geqq \delta,\ X_{k+1}^N(t) = 0) &= P(X_k^N(t) \geqq \delta,\ A^N(t) = k) \\ &= P(X_k^N(t) \geqq \delta,\ A^N(t) > k-1) - P(X_k^N(t) \geqq \delta,\ A^N(t) > k) \\ &= P(X_k^N(t) \geqq \delta) - P(X_k^N(t) \geqq \delta,\ X_{k+1}^N(t) > 0) \\ &\to P(X_k(t) \geqq \delta) - P(X_k(t) \geqq \delta,\ X_{k+1}(t) > 0) \\ &= P(X_k(t) \geqq \delta,\ D_\infty(t) = k) \\ &= P(D_\infty(t) = k) P(X_k(t) \geqq \delta \mid D_\infty(t) = k),\end{aligned} \tag{40}$$

where $X_k(t)$ denotes the kth component of $X(t)$. Also, as a consequence of (35),

$$\lim_{N \to \infty} P(Q_K^N \geqq i) = P(Q_K \geqq i) \tag{41}$$

where Q_K is the number of new classes formed in $R_K(\cdot)$, the K-coalescent with ages.

Recall (Donnelly and Tavaré (1986), Equation (5.1)) that Q_K has the representation $Q_K = \sum_{j=1}^{K} \chi_j$ where the χ_j are independent random variables with $P(\chi_j = 1) = 1 - P(\chi_j = 0) = \theta/(j + \theta - 1)$. The Borel–Cantelli lemma ensures that $Q_K \to \infty$ in probability as $K \to \infty$, so choose and fix K large enough to ensure

$$P(Q_K \geqq i) > (1 - \varepsilon)^{\frac{1}{3}}. \tag{42}$$

Also $D_\infty(t) \to \infty$ a.s. as $t \to 0$, so choose and fix $t > 0$ (depending on K) to ensure that

$$P(D_\infty(t) \geqq K) > (1 - \varepsilon)^{\frac{1}{3}} + (1 - (1 - \varepsilon)^{\frac{1}{3}})/2. \tag{43}$$

But for $t > 0$, $E(D_\infty(t)) < \infty$, so (for example by Markov's inequality), choose and fix L (depending on t) so that

$$P(D_\infty(t) > L) < (1 - (1 - \varepsilon)^{\frac{1}{3}})/2. \tag{44}$$

https://doi.org/10.2307/1427746 Published online by Cambridge University Press

Equations (43) and (44) then guarantee that

$$P(K \leqq D_\infty(t) \leqq L) > (1-\varepsilon)^{\frac{1}{3}}. \tag{45}$$

Finally note that $P(X_L(t) = 0 \mid D_\infty(t) = L) = 0$, so

$$\lim_{n\to\infty} P\left(X_L(t) > \frac{1}{n} \,\middle|\, D_\infty(t) = L\right) = P(X_L(t) > 0 \mid D_\infty(t) = L) = 1$$

and we may choose $\delta > 0$ (depending on L) with

$$P(X_L(t) \geqq \delta \mid D_\infty(t) = L) > (1-\varepsilon)^{\frac{1}{3}}. \tag{46}$$

Observe that for fixed δ, $P(X_k(t) \geqq \delta \mid D_\infty(t) = k)$ is non-increasing in k. Thus in view of (39)–(46), for the given ε and the above choice of K, t, L, and δ, we can write

$$\begin{aligned}\limsup_{N\to\infty} \lambda_N(\{x_i \geqq \delta\}) &\geqq P(Q_K \geqq i) \sum_{k=K}^{L} P(D_\infty(t) = k) P(X_k(t) \geqq \delta \mid D_\infty(t) = k) \\ &\geqq P(Q_K \geqq i) P(X_L(t) \geqq \delta \mid D_\infty(t) = L) P(K \leqq D_\infty(t) \leqq L) \\ &> 1 - \varepsilon,\end{aligned}$$

and the result follows.

Acknowledgements

It is a pleasure to acknowledge a number of helpful discussions with Bob Keener concerning the characterization results of Section 3.2. Part of the work was completed while the first author was visiting the Department of Mathematics at the University of Utah. Their warm hospitality is gratefully acknowledged.

References

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

CANNINGS, C. (1974) The latent roots of certain Markov chains arising in genetics: a new approach. 1. Haploid models. *Adv. Appl. Prob.* **6,** 260–290.

DAVIES, E. B. AND VINCENT-SMITH G. F. (1968) Tensor products, infinite products and projective limits of Choquet simplexes. *Math. Scand.* **22,** 145–164.

DONNELLY, P. (1986) Partition structures, Pólya urns, the Ewens sampling formula, and the ages of alleles, *Theoret. Popn Biol.* **30,** 271–288.

DONNELLY, P. (1991) Weak convergence to a Markov chain with an entrance boundary: ancestral processes in population genetics. *Ann. Prob.*

DONNELLY, P. AND TAVARÉ, S. (1986) The ages of alleles and a coalescent. *Adv. Appl. Prob.* **18,** 1–19.

DONNELLY, P. AND TAVARÉ, S. (1987) The population genealogy of the infinitely-many neutral alleles model. *J. Math. Biol.* **25,** 381–391.

ETHIER, S. N. AND KURTZ, T. G. (1986) *Markov Processes—Characterization and Convergence.* Wiley, New York.

https://doi.org/10.2307/1427746 Published online by Cambridge University Press

EWENS, W. J. (1972) The sampling theory of selectively neutral alleles. *Theoret. Popn Biol.* **3,** 87–112.

EWENS, W. J. (1990) Population genetics theory—the past and the future. In *Mathematical and Statistical Developments in Evolutionary Theory,* ed. S. Lessard, Kluwer Academic Publishers, Dordrecht, pp. 177–228.

KINGMAN, J. F. C. (1977) The population structure associated with the Ewens sampling formula. *Theoret. Popn Biol.* **11,** 274–283.

KINGMAN, J. F. C. (1978a) Random partitions in population genetics. *Proc. R. Soc. London* A **361,** 1–20.

KINGMAN, J. F. C. (1978b) The representation of partition structures. *J. London Math. Soc.* (2) **18,** 374–380.

KINGMAN, J. F. C. (1982a) On the genealogy of large populations. *J. Appl. Prob.* **19**A, 27–43.

KINGMAN, J. F. C. (1982b) The coalescent. *Stoch. Proc. Appl.* **13,** 235–248.

KINGMAN, J. F. C. (1982c) Exchangeability and the evolution of large populations. In *Exchangeability in Probability and Statistics,* ed. G. Koch and F. Spizzichino, North-Holland, Amsterdam, pp. 97–112.

LAURITZEN, S. L. (1974) Sufficiency, prediction, and extreme models. *Scand. J. Statist.* **1,** 128–134.

TAVARÉ, S. (1984) Line-of-descent and genealogical processes and their applications in population genetics models. *Theoret. Popn Biol.* **26,** 119–164.

WATTERSON, G. A. (1984) Lines of descent and the coalescent. *Theoret. Popn Biol.* **26,** 77–92.

https://doi.org/10.2307/1427746 Published online by Cambridge University Press