J. Math. Biology (1984) 20: 91–94

Journal of
Mathematical
Biology

Pólya-like urns and the Ewens' sampling formula

Fred M. Hoppe

Departments of Mathematics and Statistics, The University of Michigan, Ann Arbor, MI 48109, USA

Abstract. A Markov process of partitions of the natural numbers is constructed by defining a Pólya-like urn model. The marginal distributions of this process are the Ewens' sampling distributions of population genetics.

Key words: Ewens' sampling formula — Random partitions — population genetics — Pólya urn

1. Introduction

This paper is motivated by Ewens' [1, Eq. (18)], where it is shown that if a random sample of genes, each of which can be in one of many allelic states, is drawn from an evolving population which has come to an equilibrium, then the probability that the $(j+1)$th gene is of a type previously drawn is $j/(\theta+j)$ where $\theta>0$ is some constant. This behaviour mimics the conditional probabilities of drawings in a Pólya urn and it is this connection which we explore.

Consider then a process $\{X_n\}$ generated by sampling from an urn containing one black ball and various numbers of coloured (non-black) or labelled balls. The black ball has mass $\theta>0$ and every other ball has mass one. At time n a ball is selected at random (that is in proportion to its mass) from the urn. If labelled it is returned together with one additional ball of the same colour, but if black it is returned together with one additional ball of a previously unobserved colour. For definiteness we will use the natural numbers for labelling the colours and choose them sequentially as the need arises. The random variable X_n is then the label of the additional ball returned after the nth drawing. This describes the Pólya-like process. Originally there is only one ball in the urn, the black one. Thus $X_1=1$, $X_2=1$ or 2, $X_3=1$, 2 or 3, etc.

If we define the quantity K as the random number of distinct labels present then at time n the urn will contain a total of n balls of which n_i $(1\leq i\leq K)$ are numbered i and its contents may be identified with a configuration of n tokens placed in K cells. (The black ball is ignored in describing the urn configuration since it is always present and merely a device for generating new labels, that is introducing mutations.)

Research supported in part by NSF Grant MCS-8108689

This model being motivated by a genetic application in which no physical significance may be attached to the actual labels requires that all biologically meaningful information be contained in the unordered occupancy numbers $\{n_1, n_2, \ldots, n_K\}$. It is thus of interest to ascertain their distribution. To this end we recall the notion of a partition [3].

A partition of a positive integer n is a representation of n as a sum of positive integers. A set of occupancy numbers $\{n_1, \ldots, n_K\}$ where $n_1 + \cdots + n_K = n$ thus defines a partition of n. For each $1 \leq i \leq n$, let $a_i = \#$ of times the integer i appears in $\{n_1, \ldots, n_K\}$. The vector $a = (a_1, \ldots, a_n)$ is called the allelic partition corresponding to $\{n_1, \ldots, n_K\}$. It is a convenient description in genetic applications where the labels represent distinct allelic forms of a gene and a_i the number of alleles with i representatives. Observe that the sequence of draws $\{X_k\}_{k=1}^n$ results in a random partition denoted by Π_n.

2. Theorem

$\{\Pi_n\}$ is a Markov process with marginal distribution

$$P[\Pi_n = a] = \frac{n!}{[\theta]^n} \prod_{i=1}^{n} \frac{\theta^{a_i}}{i^{a_i} a_i!} \tag{1}$$

where $[\theta]^n = \theta(\theta+1)\cdots(\theta+n-1)$ is the ascending factorial and $a = (a_1, \ldots, a_n)$ with $\sum i a_i = n$.

Proof: Fix a partition $a = (a_1, \ldots, a_n)$ resulting from occupancy numbers $\{n_1, \ldots, n_K\}$ and consider a possible sample path $\{X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\}$. Clearly

$$P[X_1 = x_1, \ldots, X_n = x_n] = \frac{\theta^K \prod_{j=1}^{K} (n_j - 1)!}{[\theta]^n}. \tag{2}$$

The factor θ^K reflects the selection K times of the black ball following which each new label must be selected an additional $(n_j - 1)$ times. The denominator $[\theta]^n = \theta(\theta+1)\cdots(\theta+n-1)$ is the product of the successive masses of the balls in the urn on each of the first n drawings.

It remains to count the number of such sample paths. This obviously equals the number of distinguishable permutations of n objects which are divided into types $1, 2, \ldots, K$ subject to the following constraints:

(i) The first object of type 1 precedes the first object of type 2 which precedes the first object of type 3 and so on;

(ii) The numbers of each type are not fixed; rather there are n_1 of some type, n_2 of another, ... and n_K of a last where $n_1 + \cdots + n_K = n$.

For example if the unordered occupancy numbers are $\{1, 2, 1, 1\}$ determining the partition $a = (3, 1, 0, 0, 0)$ then the sequence of urn drawings 12234 is a permissible sample path while 13224 is not.

Introduce the following notation. Arrange the occupancy numbers in their decreasing order statistics $n_1 \geq n_2 \geq \cdots \geq n_K$. Let p equal the number of distinct integers in the set $\{n_1, \ldots, n_K\}$ and define α_1 to be the number of indices i such

that $n_i = n_1$, α_2 the number of indices i such that $n_i = n_{\alpha_1+1}$ and so on, and finally α_p the number of indices i such that $n_i = n_K$ (see [2]).

Imposing condition (ii) on the sample paths observe that there are $K!/\prod_{i=1}^{p} \alpha_i!$ distinguishable ways of distributing the numbers $\{n_1, \ldots, n_K\}$ among the K types and for each such way there are $n!/\prod_{i=1}^{K} n_i!$ permutations of labels agreeing with the occupancy numbers, resulting in a total of

$$\frac{K!n!}{\prod_{i=1}^{p} \alpha_i! \prod_{j=1}^{K} n_j!} \tag{3}$$

permutations fulfilling condition (ii).

Not all among these will obey condition (i). Separate the permutations into disjoint classes determined by the order of first appearance of the digits $\{1, 2, \ldots, K\}$. There are $K!$ such disjoint classes each having, by symmetry, the same cardinality, and only one satisfying (1). If we multiply (2) by (3) and divide by $K!$ we will obtain

$$\frac{n!}{[\theta]^n} \frac{\theta^K}{\prod_{j=1}^{K} n_j \prod_{i=1}^{p} \alpha_i!}$$

which is just another way of expressing (1). Hence the marginal distributions are just the Ewens' distributions for different sample sizes n. The Markovian assertion is immediate.

Remarks: Equation (1) is known as Ewens' sampling formula in population genetics [1]. It arises in various models (see Kingman [3] who has identified three broad features linking the models) in all of which a population with genetic material is evolving through reproduction and mutation and (1) is essentially the limit distribution of the genetic content in a random sample taken from the population. This derivation exhibits Ewens' formula as the marginal distribution of a Markov process of partitions, the main novelty thus being that there is no limit involved in obtaining the distribution.

3. Connection with combinatorics

As an example of the usefulness of the above we explain the appearance [3] of the combinatorial component $n!/\prod_{i=1}^{n} i^{a_i} a_i!$. This function of a is known to equal the number of permutations of n objects whose cycle decomposition has precisely a_i cycles of length i, and we now give an alternate proof of the preceding theorem, based on cyclic permutations, to count the number of sample paths. The approach is reminiscent of the well-known duality between urn models and occupancy problems.

Consider an object set $\mathbf{O} = \{A_1, \ldots, A_n\}$ and σ a permutation on $\mathbf{O}$. Express σ as a product of cyclic permutations, say K cycles. With each cycle associate a label coded by an integer in $\{1, 2, \ldots, K\}$. Each element in $\mathbf{O}$ appears in one and only one cycle and thus acquires a label. Now σ also has a permutation representation as a member of S_n the symmetric group on n objects. In this representation replace each element of the object set by its label.

Here is an example with $K=3$. Let $\mathbf{O}=\{A, B, C, D, E, F, G\}$ and consider the permutation $\sigma(ABCDEFG)=(CEDABGF)$. Its cycle representation is $(ACD)(BE)(FG)$. Consider then the labelling $(ACD)\leftrightarrow 1$, $(BE)\leftrightarrow 2$, $(FG)\leftrightarrow 3$. Using this labelling we find that $(CEDABGF)$ becomes (1211233). Similarly with the labelling $(ACD)\leftrightarrow 1$, $(BE)\leftrightarrow 1$, $(FG)\leftrightarrow 3$, $(CEDABGF)$ becomes (2122133).

There are $K!$ labellings possible but in only one of these will it be the case that in the permutation representation the first appearance of 1 will precede the first appearance of 2 which will precede the first appearance of 3 and so on. Choose this labelling to define a mapping T from permutations on n objects to permutations of K integers satisfying the previous condition (i). Now this mapping is not $1-1$. For instance with the above notation the permutation $\sigma'(ABCDEFG)=(DEACBGF)$ has the cycle representation $(ADC)(BE)(FG)$ and the labelling $(ADC)\leftrightarrow 1$, $(BE)\leftrightarrow 2$, $(FG)\leftrightarrow 3$ transforms $(DEACBGF)$ into (1211233) which is the same as in the first example of the previous paragraph.

Two permutations will map into the same sequence of K integers fulfilling (i) if and only if their cycle representations contain the same groups of symbols. Fixing one member of a cycle uniquely determines the cycle so that if a cycle contains n_j elements there are $(n_j-1)!$ distinct cycles with the same elements. Hence the mapping T has multiplicity $\prod_{j=1}^{K}(n_j-1)!$ on those permutations whose cycle representation has cycle numbers $\{n_1, \ldots, n_K\}$.

To each permutation σ there corresponds a partition $a=(a_1, \ldots, a_n)$ of n defined by $a_i \equiv a_i(\sigma)=$ the number of cycles of length i in the cyclic representation of σ and the number of permutations whose cycle decomposition determines a fixed partition $(a_1, \ldots, a_n)$ is given by $n!/\prod_{i=1}^{n} i^{a_i} a_i!$. Hence the number of distinct sample paths satisfying (i) and (ii) is $n!/(\prod_{i=1}^{n} i^{a_i} a_i!)(\prod_{j=1}^{K}(n_j-1)!)$. Multiplication of this expression by (2) will give (1) again.

References

1. Ewens, W. J.: The sampling theory of selectively neutral alleles. Theoret Population Biology **3**, 87–112 (1972)
2. Karlin, S., McGregor, J.: Addendum to a paper of W. Ewens. Theoret Population Biology **3**, 113–116 (1972)
3. Kingman, J. F. C.: Mathematics of Genetic Diversity, SIAM, Washington (1980)

Received January 3/Revised May 1, 1984