Statistics, Probability and Game Theory
IMS Lecture Notes – Monograph Series (1996) Volume 30

SOME DEVELOPMENTS OF THE BLACKWELL-MACQUEEN URN SCHEME

JIM PITMAN
University of California

Abstract

The Blackwell-MacQueen description of sampling from a Dirichlet random distribution on an abstract space is reviewed, and extended to a general family of random discrete distributions. Results are obtained by application of Kingman's theory of partition structures.

1 Introduction

Blackwell and MacQueen [10] described the construction of a Dirichlet prior distribution by a generalization of Pólya's urn scheme. While the notion of a random discrete probability measure governed by a Dirichlet distribution was first developed in the setting of Bayesian statistics [30, 26, 27, 28], this idea has applications in other fields. The distribution of the ranked masses of atoms in a Dirichlet distribution, called the *Poisson-Dirichlet* (PD) *distribution* [45], appears as an asymptotic distribution in number theory [14, 8, 67, 16], combinatorics [65, 68, 69, 34], and population genetics [70, 24]. Though the finite dimensional distributions of the PD distribution are difficult to describe explicitly, there are some remarkably simple formulae involving this distribution, most notably the Ewens sampling formula [23, 25]. Antoniak [3] derived the Ewens sampling formula from the Blackwell-MacQueen description of sampling from a Dirichlet prior. Hoppe [35, 37] used the urn scheme to derive the simple form of the *size-biased random permutation* of the PD distribution, which Ewens [24] termed the GEM distribution, after Griffiths, Engen and McCloskey, who contributed to its development and application in the fields of genetics and ecology. Dirichlet random measures and the PD and GEM distributions appear also as the stationary distributions of measure-valued diffusions derived from population genetics models [19, 20, 21, 22].

Section 2 of this paper reviews some basic results involving Dirichlet distributions and the PD and GEM distributions. Section 3 shows how many of these results extend to a more general Bayesian model for sampling from a random discrete distribution. This involves Kingman's theory of partition structures [46] as developed in [1, 55]. The general model is illustrated by a two-parameter model for species sampling, first proposed by Engen [18],