Ann. Inst. Statist. Math.
Vol. 45, No. 3, 453–458 (1993)

A PÓLYA URN MODEL WITH A CONTINUUM OF COLORS

HAJIME YAMATO

Department of Mathematics, Faculty of Science, Kagoshima University, 1-21-35 Kourimoto, Kagoshima 890, Japan

(Received October 7, 1991; revised September 14, 1992)

Abstract. For a Pólya urn model with a continuum of colors introduced by Blackwell and MacQueen ((1973), *Ann. Statist.*, **2**, 1152–1174), we show the joint distribution of colors after n draws from which several properties of the urn model are derived. The similar results hold for the case where the initial distribution of colors is invariant under a finite group of transformations.

Key words and phrases: Dirichlet process, Dirichlet invariant process, Stirling numbers of the first kind.

1. Introduction

Blackwell and MacQueen (1973) extended the Pólya urn model by allowing a continuum of colors and derived the Ferguson's Dirichlet processes as the limit $(n \to \infty)$ of the distribution of colors after n draws. This urn model may be described as follows. A color is initially chosen from a continuous probability distribution Q (on the d-dimensional Euclidean space R^d) and r balls of this color are put in an empty urn. Then, successively after n draws, with probability $M/(M+nr)$ a color is chosen from the probability distribution Q and r balls of this color are put in the urn, or with probability $nr/(M+nr)$ a ball is drawn from the urn and returned to it with r balls of the same color. Let $X_1, X_2, \dots, X_n, \dots$ be the sequence of chosen colors. Let $C(m_1, \dots, m_n : n)$ denote the set of the first n trials in which m_i colors appear i times, $i = 1, 2, \dots, n$ so that $\sum_{i=1}^{n} im_i = n$. Given $(X_1, \dots, X_n) \in C(m_1, \dots, m_n : n)$, let $X_{i,1}, \dots, X_{i,m_i}$ be the m_i colors appearing i times, $i = 1, 2, \dots, n$. Note that when Q is a discrete distribution on the finite set this model reduces to the usual Pólya urn model.

In the present paper the probability $\Pr\{X_{ij} \in A_{ij} (i = 1, \dots, n, j = 1, \dots, m_i), (X_1, \dots, X_n) \in C(m_1, \dots, m_n : n)\}$,where $A_{ij} \subseteq R^d$ is a Borel set, is inductively derived. As a corollary, (a) the probability $\Pr\{(X_1, \dots, X_n) \in C(m_1, \dots, m_n : n)\}$ is deduced, (b) the probability function of the number D_n of distinct colors among $X_1, \dots, X_n$ is obtained in terms of the absolute (unsigned) Stirling numbers of the first kind and (c) the new observed colors $Y_1, Y_2, \dots$ are shown to be independent and identically distributed random variables. The estimation of the parameters M and Q is discussed when they are unknown. These are shown in Section 2.

Finally the case with Q a probability distribution invariant under a finite group of transformations is discussed in Section 3.

2. A Pólya urn model

The sequence of colors $X_1, X_2, \ldots, X_n, \ldots$ in the Pólya urn model stated in Section 1 is formalized as follows: for any Borel set $A \subseteq R^d$

$$\text{(A)} \qquad \Pr(X_1 \in A) = Q(A)$$

and

$$\text{(B)} \qquad \Pr(X_{n+1} \in A \mid X_1 = x_1, \ldots, X_n = x_n) = \left[MQ(A) + r \sum_{i=1}^{n} \delta_{X_i}(A) \right] / (M + nr),$$

where Q is a continuous probability distribution on R^d, $\delta_x(A) = 1$ if $x \in A$ and $= 0$ otherwise, M is a positive constant and r is a positive integer. By introducing parameter $M^* = M/r$, the condition (B) may be written as

$$\text{(B}^*\text{)} \qquad \Pr(X_{n+1} \in A \mid X_1 = x_1, \ldots, X_n = x_n) = \left[M^* Q(A) + \sum_{i=1}^{n} \delta_{X_i}(A) \right] / (M^* + n).$$

Hence we assume $r = 1$ in the condition (B) in the sequel. We have the following theorem.

THEOREM 2.1. *Under the Pólya urn model with a continuum of colors described by* (A) *and* (B) *with* $r = 1$, *we have for any Borel sets* $A_{ij} \subseteq R^d$ $(i = 1, \ldots, n,\ j = 1, \ldots, m_i)$,

$$\text{(2.1)} \qquad \begin{aligned} &\Pr\{X_{ij} \in A_{ij} (i = 1, \ldots, n, j = 1, \ldots, m_i), \\ &\qquad (X_1, \ldots, X_n) \in C(m_1, \ldots, m_n : n)\} \\ &= \frac{n!}{\prod_{i=1}^{n} m_i! i^{m_i}} \cdot \frac{M^{\sum m_i}}{M^{[n]}} \prod_{i=1}^{n} \prod_{j=1}^{m_i} Q(A_{ij}), \end{aligned}$$

where $M^{[n]} = M(M+1) \cdots (M+n-1)$.

PROOF. Let X_{ij}^*, $i = 1, \ldots, n+1$ and $j = 1, \ldots, m_i$, be the distinct colors among $X_1, \ldots, X_{n+1}$ $(\in C(m_1, \ldots, m_{n+1} : n+1))$, where $\sum_{i=1}^{n+1} i m_i = n+1$. From the condition (B) with $r = 1$ and the continuity of Q, we have for any Borel set $A \subseteq R^d$, $\Pr\{X_{n+1}$ is new and $\in A \mid X_1 = x_1, \ldots, X_n = x_n\} = MQ(A)/(M+n)$, which is independent of the previous observations. Then we have the recurrence relation,

$$
\begin{aligned}
&\Pr\{X^*_{ij} \in A_{ij}(i = 1, \ldots, n+1, j = 1, \ldots, m_i),\\
&\qquad (X_1, \ldots, X_{n+1}) \in C(m_1, \ldots, m_{n+1} : n+1)\}\\
&= \frac{M}{M+n} \Pr\{X_{n+1} \text{ is new and } \in A_{1,m_1}\}\\
&\quad \cdot \Pr\{X_{1j} \in A_{1j}(j = 1, \ldots, m_1 - 1), X_{ij} \in A_{ij}\\
&\qquad\quad (i = 2, \ldots, n, j = 1, \ldots, m_i), (X_1, \ldots, X_n) \in C(m_1 - 1, \ldots, m_n : n)\}\\
&\quad + \sum_{r=1}^{n} \sum_{l=1}^{m_r+1} \frac{r}{M+n}\\
&\quad \cdot \Pr\{X_{ij} \in A_{ij}(i = 1, \ldots, n(\neq r, r+1), j = 1, \ldots, m_i),\\
&\qquad X_{rj} \in A_{rj}(j = 1, \ldots, l-1), X_{r,j+1} \in A_{rj}(j = l, \ldots, m_r),\\
&\qquad X_{r+1,j} \in A_{r+1,j}(j = 1, \ldots, m_{r+1} - 1), X_{rl} \in A_{r+1,m_{r+1}}\\
&\qquad (X_1, \ldots, X_n) \in C(m_1, \ldots, m_r + 1, m_{r+1} - 1, \ldots, m_n : n)\}.
\end{aligned}
$$

Using this relation the probability (2.1) is proved by induction. □

If we take $A_{ij} = R^d$ $(i = 1, \ldots, n, j = 1, \ldots, m_i)$, then from Theorem 2.1 we have the following which is essentially equivalent to Proposition 3 of Antoniak (1974).

COROLLARY 2.1. *For the Pólya urn model with a continuum of colors, we have*

$$\Pr\{(X_1, \ldots, X_n) \in C(m_1, \ldots, m_n : n)\} = \frac{n!}{\prod_{i=1}^{n} m_i! i^{m_i}} \cdot \frac{M^{\sum m_i}}{M^{[n]}}. \tag{2.2}$$

Let D_n denote the number of distinct colors among $X_1, \ldots, X_n$. For $(X_1, \ldots, X_n) \in C(m_1, \ldots, m_n : n)$, we have $D_n = \sum_{i=1}^{n} m_i$. The sum of the events $\{(X_1, \ldots, X_n) \in C(m_1, \ldots, m_n : n)\}$ over $(m_1, \ldots, m_n)$ satisfying $m_1 + 2m_2 + \cdots + nm_n = n$ and $m_1 + \cdots + m_n = k$ is equal to the event $\{D_n = k\}$. The sum of $n!/[\prod m_i! i^{m_i}]$ over $(m_1, \ldots, m_n)$ satisfying the same conditions is equal to the absolute Stirling numbers of the first kind $|s(n, k)|$ (see for example Comtet (1974)). Thus, from (2.2) we have the following, which is essentially equivalent to the distribution of Z_n given by Antoniak ((1974), p. 1161).

COROLLARY 2.2. *Under the Pólya urn model with a continuum of colors, we have for $k = 1, \ldots, n$,*

$$\Pr\{D_n = k\} = |s(n, k)| \frac{M^k}{M^{[n]}}, \tag{2.3}$$

where $s(n, k)$ is Stirling numbers of the first kind.

By dividing (2.1) by (2.2) we have easily the following.

COROLLARY 2.3. *Under the Pólya urn model with a continuum of colors, we have for any Borel sets* $A_{ij} \subseteq R^d$ $(i = 1, \ldots, n,\ j = 1, \ldots, m_i)$,

$$\Pr\{X_{ij} \in A_{ij}(i = 1, \ldots, n, j = 1, \ldots, m_i) \mid (X_1, \ldots, X_n) \in C(m_1, \ldots, m_n : n)\} = \prod_{i=1}^{n} \prod_{j=1}^{m_i} Q(A_{ij}). \tag{2.4}$$

Let $Y_1, Y_2, \ldots$ be the sequence of new colors in the Pólya urn model with parameters M and Q. Given $(X_1, \ldots, X_n) \in C(m_1, \ldots, m_n : n)$ with $\sum_{i=1}^{n} m_i = k$, $Y_1, \ldots, Y_k$ are the distinct colors among $X_1, \ldots, X_n$. Then we have the following.

COROLLARY 2.4. *Under the Pólya urn model with a continuum of colors, we have for any Borel sets* $A_j \subseteq R^d$ $(j = 1, \ldots, k$ *and* $k = \sum_{i=1}^{n} m_i)$,

$$\Pr\{Y_j \in A_j(j = 1, \ldots, k) \mid (X_1, \ldots, X_n) \in C(m_1, \ldots, m_n : n)\} = \prod_{j=1}^{k} Q(A_j). \tag{2.5}$$

PROOF. The equation (2.4) implies that X_{ij}, $i = 1, \ldots, n$ and $j = 1, \ldots, m_i$ are independent and so exchangeable, given $(X_1, \ldots, X_n) \in C(m_1, \ldots, m_n : n)$. Thus $(Y_1, \ldots, Y_k)$ are one of all permutations of $X_{11}, \ldots, X_{1m_1}, X_{21}, X_{22}, \ldots, X_{2m_2}, X_{31}, X_{32}, \ldots$ with probability $1/k!$, given $(X_1, \ldots, X_n) \in C(m_1, \ldots, m_n : n)$. Therefore, by (2.4) we have the equation (2.5).

Since the probability (2.5) is the same for any $(m_1, \ldots, m_n)$ satisfying $k = \sum_{i=1}^{n} m_i$, we have following Corollary 2.5. Corollaries 2.5 and 2.6 are essentially equivalent to Theorems 2.5 and 2.7 of Korwar and Hollander (1973), respectively.

COROLLARY 2.5. *Under the Pólya urn model with a continuum of colors, we have for any Borel sets* A_j $(j = 1, \ldots, k)$,

$$\Pr\{Y_j \in A_j(j = 1, \ldots, k) \mid D_n = k\} = \prod_{j=1}^{k} Q(A_j). \tag{2.6}$$

COROLLARY 2.6. *For the Pólya urn model with a continuum of colors,* $Y_1, Y_2, \ldots$ *are independent and identically distributed with the distribution* Q.

PROOF. From (2.3) and (2.6) we have for any Borel sets A_j $(j = 1, \ldots, k)$,

$$\begin{aligned}
\Pr\{Y_j \in A_j(j = 1, \ldots, k)\}
&= \Pr\{Y_j \in A_j(j = 1, \ldots, k), D_k = k\} \\
&\quad + \sum_{n=k}^{\infty} \Pr\{Y_j \in A_j(j = 1, \ldots, k), D_n = k-1, X_{n+1} \text{ is new and } \in A_k\} \\
&= \prod_{j=1}^{k} Q(A_j) M^k \sum_{n=k-1}^{\infty} \frac{|s(n, k-1)|}{M^{[n+1]}} = \prod_{j=1}^{k} Q(A_j),
\end{aligned}$$

because of the well-known property of Stirling numbers of the first kind, $\sum_{n=k}^{\infty} |s(n,k)|/M^{[n+1]} = 1/M^{k+1}$ (see for example, p. 2545 of Charalambides and Singh (1988)). Since we have $\Pr\{Y_i \in A_i (i = 1, \ldots, k)\} = \prod_{i=1}^{k} Q(A_i)$ for any integer k (≥ 2), $Y_1, Y_2, \ldots$ are independent and identically distributed with Q. □

As an application of the above results, we consider the estimation problem of parameters M and Q when they are unknown. From (2.1) and (2.3), the conditional probability $\Pr\{X_{ij} \in A_{ij} (i = 1, \ldots, n, j = 1, \ldots, m_i), (X_1, \ldots, X_n) \in C(m_1, \ldots, m_n : n) \mid D_n = k\}$ does not depend on parameter M. Thus D_n is a sufficient statistic for M. It is easily shown that D_n is complete. The maximum likelihood estimator $\hat{M}$ of M is given by maximizing $L(M) = M^{D(n)}/M^{[n]}$ from (2.3). Especially if $D(n) = 1$, $L(M)$ is monotone decreasing and $\hat{M} = 0$, and if $D(n) = n$, $L(M)$ monotone increasing and $\hat{M} = \infty$. Otherwise, put $l(M) = \log L(M)$, then we have $l'(M) = \{D(n) - h(M)\}/M$ where $h(M) = 1 + M/(M+1) + \cdots + M/(M+n-1)$. $h(M)$ ($M > 0$) is a monotone increasing function taking values between 1 and n and $l'(M) = 0$ has a unique solution $\hat{M}$ which is the maximum likelihood estimator. Generally $\hat{M}$ must be evaluated numerically (see for example Sibuya (1991)). From (2.4), given $(X_1, \ldots, X_n) \in C(m_1, \ldots, m_n : n)$, the distinct colors X_{ij}, $i = 1, \ldots, n$ and $j = 1, \ldots, m_i$, are independent and identically distributed with the continuous distribution Q. Since from (2.2), $(X_1, \ldots, X_n) \in C(m_1, \ldots, m_n : n)$ does not depend on Q, the empirical distribution function based on the distinct colors $X_{ij}, i = 1, \ldots, n$ and $j = 1, \ldots, m_i$, is a best estimator of Q.

3. A Pólya urn model with an invariant distribution of colors

We consider the Pólya urn model with a continuum of colors under an additional condition that Q is invariant under the finite group of transformations on R^d, $G = \{g_1, \ldots, g_u\}$. We assume the condition (A) given in Section 2 and the following condition (B′) instead of the condition (B) given in Section 2:

$$\text{(B}'\text{)}\ \Pr(X_{n+1} \in A \mid X_1 = x_1, \ldots, X_n = x_n) = \left[MQ(A) + \sum_{i=1}^{n} \delta^*_{X_i}(A) \right] /(M+n),$$

where $\delta^*_x(A) = (1/u) \sum_{j=1}^{u} \delta_{g_j x}(A)$.

Let $O(x)$ be the orbit of x for the group of transformations G, that is, $O(x) = \{g_1 x, \ldots, g_u x\}$. Let $X_1, \ldots, X_n$ be the sequence of chosen colors. Let $C^*(m_1, \ldots, m_n : n)$ denote the set of the first n trials in which m_i orbits of colors appear i times, $i = 1, \ldots, n$. Given $(X_1, \ldots, X_n) \in C^*(m_1, \ldots, m_n : n)$, let $X_{i,1}, \ldots, X_{i,m_i}$ be the m_i colors whose orbits appear i times, $i = 1, \ldots, n$. Then we have the following theorem similar to Theorem 2.1.

THEOREM 3.1. *Under the Pólya urn model having a continuum of colors described by* (A) *and* (B′) *with G-invariant distribution Q, we have for any G-invariant Borel sets $B_{ij} \subseteq R^d$ ($i = 1, \ldots, n$, $j = 1, \ldots, m_i$),*

$$\Pr\{X_{ij} \in B_{ij} (i = 1, \ldots, n, j = 1, \ldots, m_i), (X_1, \ldots, X_n) \in C^*(m_1, \ldots, m_n : n)\}$$

$$= \frac{n!}{\prod_{i=1}^{n} m_i! i^{m_i}} \cdot \frac{M^{\sum m_i}}{M^{[n]}} \prod_{i=1}^{n} \prod_{j=1}^{m_i} Q(B_{ij}).$$

Further we have the propositions similar to Corollaries 2.1–2.3 and 2.4–2.6. These results gives also a characterization of a sample from a distribution having Dirichlet invariant process (see Yamato (1987)). For the Dirichlet invariant process see for example Dalal (1979).

Acknowledgements

The author would like to express his deep gratitude to Prof. Masaaki Sibuya for his valuable suggestions. He also thanks the referees for their rigorous and valuable suggestions.

References

Antoniak, C. A. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems, *Ann. Statist.*, **2**, 1152–1174.

Blackwell, D. and MacQueen, J. (1973). Ferguson distribution via Pólya urn schemes, *Ann. Statist.*, **1**, 353–355.

Charalambides, Ch. A. and Singh, J. (1988). A review of the Stirling numbers, their generalization and a statistical applications, *Comm. Statist. Theory Methods*, **17**, 2533–2593.

Comtet, L. (1974). *Advanced Combinatorics*, Reidel, Dordrecht.

Dalal, S. R. (1979). Dirichlet invariant processes and applications to nonparametric estimation of symmetric distribution functions, *Stochastic Process. Appl.*, **9**, 99–107.

Korwar, R. M. and Hollander, M. (1973). Contributions to the theory of Dirichlet processes, *Ann. Probab.*, **1**, 705–711.

Sibuya, M. (1991). A cluster-number distribution and its application to the analysis of homonyms, *Japanese Journal of Applied Statistics*, **20**, 139–153 (in Japanese).

Yamato, H. (1987). On samples from distributions chosen from a Dirichlet invariant process, *Bull. Inform. Cybernet.*, **22**, 199–207.