ON CHARACTERIZING THE PÓLYA DISTRIBUTION

HÉCTOR M. RAMOS¹, DAVID ALMORZA AND JUAN A. GARCÍA–RAMOS

Abstract. In this paper two characterizations of the Pólya distribution are obtained when its contagion parameter is negative. One of them is based on mixtures and the other one is obtained by characterizing a subfamily of the discrete Pearson system.

Mathematics Subject Classification. 60E05, 60E99.

Received September 14, 2001. Revised March 5, 2002.

1. Introduction

The Pólya distribution was derived by Eggenberger and Pólya [2]. There are some basic references in Jordan [10], Eggenberger and Pólya [3], Pólya [18], Feller [4], Friedman [5], Hald [6], Bosch [1], Patil and Joshi [16], Ord [13], Janardan and Schaeffer [8], Johnson and Kotz [9], Janardan [7], Panaretos and Xekalaki [14,15] and Philippou et al. [17]. The Pólya distribution is generally presented in terms of random drawings of balls from an urn. Initially, it is assumed that there are $N$ balls in the urn, $A$ white balls and $B = N - A$ black balls. One ball is drawn at random and then replaced with $c$ $(c \in \mathbb{Z})$ additional balls of the same color. This procedure is repeated $s$ times. The total number $X$ of the white balls in the sample has the Pólya distribution $P(N, A, s, c)$. Its probability mass function (pmf) is:

$$p[i; N, A, s, c] = p_i = \binom{s}{i} \cdot \frac{A^{(i,c)} B^{(s-i,c)}}{N^{(s,c)}}$$

where $N > A > 0$, $i = 0, 1, 2, \ldots, s$. The expression $I^{(a,c)}$ is given by:

$$I^{(a,c)} = I(I + c)(I + 2c)\ldots(I + (a - 1)c); I^{(0,c)} = 1.$$  

If $c < 0$ it is generally assumed that $(-c)(s - 1) \leq \min(A, B)$.

Note that when $\frac{A}{c}$ and $\frac{B}{c}$ are integers $(c < 0)$, then the support

$$\max \left(0, s + \frac{B}{c}\right) \leq i \leq \min \left(s, \frac{-A}{c}\right)$$

can be assumed.

Keywords and phrases: Pólya distribution, hypergeometric distribution, characterization.

¹ Departamento de Estadística e Investigación Operativa, Facultad de Ciencias Económicas y Empresariales, Universidad de Cádiz, c/Duque de Nájera 8, 11002 Cádiz, Spain; e-mail: hector.ramos@uca.es

© EDP Sciences, SMAI 2002
The constant $c$ is interpreted as a contagion parameter. If $c = -1$ the outcome is the hypergeometric distribution $H(N, A, s)$, and if $c = 0$ the outcome is the binomial distribution $B(s, A/N)$.

Although the properties of the Polya distribution are well known, there are only a few characterizations of this distribution. Janardan [7] gives a characterization based on mixtures of Polya distributions with respect to the parameter $s$ when it has a negative binomial distribution.

In this paper we will consider the Polya distribution $P(N, A, s, c)$ when $c$ is negative, and we will obtain two characterizations of this family. The first one is based on mixtures with respect to the parameter $a(a = -A/c)$, when $a$ has a Polya distribution. As a particular case, when $c = -1$, one obtains the characterization of the hypergeometric distribution given by Skibinsky [19].

The second result will be obtained by characterizing a subfamily of the discrete Pearson system. This system is formed by those distributions $F$, with pmf $\{p_i\}$ over a regular lattice $T$ of width one, which satisfy the following difference equation

$$
\triangle p_{i-1} = \frac{a - i}{b_0 + b_1 i + b_2 i(i - 1)} \cdot p_{i-1}, \quad i \in T.
$$

(1.1)

A complete classification of the Pearson system and a detailed study of this family can be found in Ord [12,13].

Ollero and Ramos [11] defined the subclass $P_H$ of the discrete Pearson system formed by those distributions with finite support $T = [m, M]$, $m \geq 0$, such that

1. $b_0 = 0$, $b_2 \neq 0$
2. $\frac{b_1}{b_2} + 2m \in \mathbb{R}^+$.

They show that if $F \sim P(N, A, s, c)$, $c < 0$, then $F \in P_H$ and the corresponding parameters in (1.1) are

$$
a = \frac{(A - c)(s + 1)}{N - 2c}; \quad b_0 = 0; \quad b_1 = \frac{B + c(s - 1)}{N - 2c}; \quad b_2 = \frac{-c}{N - 2c} \neq 0.
$$

(1.2)

In this paper, we complete these results, as we identify the family $P_H$ with the Polya distributions $P(N, A, s, c)$, $c < 0$.

The following definitions will be used:

**Definition 1.1.** Let $p_\theta(x)$ be a family of pmf indexed by the parameter $\theta$. If $\theta$ is a random variable and $f(\theta)$ its pmf, then $\sum p_\theta(x)$ is another pmf that will be called a mixture.

**Definition 1.2.** A family of distributions $\{p_\theta : \theta \in \Omega\}$ of a random variable $X$ indexed by the parameter $\theta$ in the set $\Omega$ is called complete if, for any function $u(x)$ independent of $\theta$, $E[u(x)] = 0$ for every $\theta \in \Omega$ implies $u(x) = 0$ for all $x$ (except possibly for a set of $x$ with probability measure zero for all $\theta \in \Omega$).

**Definition 1.3.** Let $F_1 \sim P(N_1, A_1, s_1, c_1)$, $c_1 < 0$, and $F_2 \sim P(N_2, A_2, s_2, c_2)$, $c_2 < 0$. We will say that $F_1$ is related to $F_2$ (denoted $F_1 \equiv d F_2$), if:

(i) $\frac{N_1}{N_2} = \frac{A_1}{A_2} = \frac{c_1}{c_2}$

(ii) $s_1 = s_2$.

**Note 1.1.** $R$ is an equivalence relation, and if $F_1 \equiv d F_2$, then $F_1 \equiv d F_2$.

**Note 1.2.** Let $F \sim P(N, A, s, c)$, $c < 0$. If we consider $F' \sim P\left(N', -sc, \frac{A}{c}, c\right)$, then $F \equiv d F'$.

### 2. Characterization based on mixtures

**Lemma 2.1.** The mixture of the family of Polya distributions $P(N, A, s, c)$ when $c < 0$, $N = -nc$, $A = -ac$, indexed by the parameter $a$, when $a$ has a Polya distribution $P(N', A', n, c')$, is another Polya distribution $P(N', A', s, c')$. 
Proof. The pmf of the mixture is:
\[
    p_i = \sum_j \binom{s}{i} \binom{n - j - c}{s - i} \cdot p[j; N', A', n, c'] \\
    = \sum_j \binom{s}{i} \frac{(-jc)_{(j-n)c}}{(-nc)_{(s-c)}} \cdot \frac{(n-j)_{(n-j-c')}}{N'(n-c')} \cdot A'(j, c') (N' - A')^{(n-j-c')}
\]
and since
\[
    (-Ic)^{(J, c)} = (-Ic)(-Ic + c) \cdots (-Ic + (J-1)c) \\
    = I(-c)(I-1)(-c) \cdots (I-J+1)(-c) = (-c)^J J! / (I-J)! = (-c)^J \cdot \binom{J}{I} \cdot J!,
\]
the pmf can be written as
\[
    p_i = \sum_j \binom{s}{i} \binom{n - j}{s - i} \cdot \frac{(n-j)_{(n-j-c')}}{N'(n-c')} \cdot A'(j, c') \frac{A'(N' - A')^{(n-j-c')}}{N'(n-c')}
\]
Therefore since \( I^{(a, c')} = I^{(b, c')(I + bc')/(a-b, c')} \), we have that
\[
    p_i = \binom{s}{i} \frac{A'(i, c') (N' - A')^{(s-i, c')}}{N'(s, c')} \cdot \frac{(n-j)_{(n-j-c')}}{N'(n-c')} \cdot A'(j, c') \frac{A'(N' - A')^{(n-j-c')}}{N'(n-c')}.
\]
The sum of the right side is 1 because it is the sum of all the probability mass of a Pólya distribution \( P(N' + sc', A' + ic', n - s, c') \). Hence,
\[
    p_i = \binom{s}{i} \frac{A'(i, c') (N' - A')^{(s-i, c')}}{N'(s, c')} \cdot \frac{(n-j)_{(n-j-c')}}{N'(n-c')} \cdot A'(j, c') \frac{A'(N' - A')^{(n-j-c')}}{N'(n-c')}
\]
and the distribution of the mixture is \( P(N', A', s, c') \). \(\square\)

**Theorem 2.1.** A family \( P_a \) of \( n + 1 \) probability distributions indexed by \( a = 0, 1, 2, \ldots, n \), with support \( T = [m, M] \), \( 0 \leq m < M \leq n \), is the family of Pólya distributions \( P(N, A, s, c) \) with \( c < 0, N = -nc \) and \( A = -ac \), if and only if:

(i) the mixture distribution when \( a \) has a Pólya distribution \( P(N', A', n, c) \) is the Pólya distribution \( P(N', A', s, c) \);

(ii) the family of distributions \( P_a(a = 0, 1, 2, \ldots, n) \) is independent of \( N \).

Proof. If \( P_a \) is the family of Pólya distributions \( P(N, A, s, c) \) with \( c < 0, N = -nc \) and \( A = -ac \), condition (i) is a consequence of the Lemma 2.1 and condition (ii) is trivial.

Conversely, if conditions (i) and (ii) hold, we will prove that \( P_a \) is the family of Pólya distributions \( P(N, A, s, c) \) with \( c < 0, N = -nc \) and \( A = -ac \).
(a) If \( c' \neq 0 \), by (i)

\[
\sum_{a=0}^{n} p_a(i) \cdot p[a; N', A', n, c'] = p[i; N', A', s, c'], i = 0, 1, 2, \ldots, s
\]  

(2.1)

which is linear in the unknowns \( p_0(i), p_1(i), \ldots, p_n(i) \), for every \( i \).

Since (ii) holds, we can obtain a system of \( n + 1 \) linear equations with \( n + 1 \) unknowns by considering \( N' + ac' \) with \( a = 0, 1, 2, \ldots, n \). We already proved that

\[
p_a(i) = \binom{s}{i} \cdot \frac{A^{(i,c)}(N - A)^{(s-i,c)}}{N(s,c)}
\]

is a solution of the system. Let us prove that this solution is unique. The determinant \( \Delta \) of the matrix of the coefficients of the system is:

\[
\begin{vmatrix}
\binom{n}{0} A^{(0,c)} & \binom{n}{1} A^{(1,c)} & \ldots & \binom{n}{n} A^{(n,c)} \\
D^{(n,c)} & D^{(n-1,c)} & \ldots & D^{(0,c)} \\
(D + c')^{(n,c')} & (D + c')^{(n-1,c')} & \ldots & (D + c')^{(0,c')} \\
(D + nc')^{(n,c')} & (D + nc')^{(n-1,c')} & \ldots & (D + nc')^{(0,c')} \\
\end{vmatrix}
\]

where \( D = N' - A' \).

Sufficiency will be proved if \( \Delta \neq 0 \). In order to prove that \( \Delta \neq 0 \) we can subtract to each row in the determinant from the previous one, and we can repeat this process \( n \) times. In this way, at the \( i \)-th step, we will subtract the row \( j - 1 \) from the row \( j \), for \( j = n + 1, n, \ldots, i + 1 \).

Since \( I^{(0,c)} = 1 \), and \( (D + tc')^{(i,c')} - (D + (t - 1)c')^{(i,c')} = (D + tc')^{(i-1,c')}c' \) \( (\nu = 1, 2, \ldots, n) \), the determinant can be written:

\[
\Delta = 
\begin{vmatrix}
(D^{(n,c')}) & \ldots & D^{(2,c')} & D & 1 \\
(D + c')^{(n-1,c')} & \ldots & (D + c')^{(0,c')} & 0 & 0 \\
(D + 2c')^{(n-2,c')} & \ldots & 2(c')^2 & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
(n!c') & \ldots & 0 & 0 & 0 \\
\end{vmatrix}
\]

and so every element under the diagonal of the determinant is zero; hence, \( \Delta \neq 0 \).

(b) If \( c' = 0 \) then \( P(N', A', n, c') = B(n, \theta) \) and \( P(N', A', s, c') = B(s, \theta) \) with \( \theta = A'/N' \) and (2.1) becomes

\[
\sum_{a=0}^{n} p_a(i) \cdot b[a; n, \theta] = b[i; s, \theta], i = 0, 1, 2, \ldots, s
\]  

(2.2)

and \( p_0(i), p_1(i), \ldots, p_n(i) \) are related by a linear equation for every \( i \). By Lemma 2.1 \( p_a(i) = p[i; N, A, s, c] \) is a solution of (2.2). We must prove that this is the unique solution of (2.2). If \( f_a(i) \) (independent of \( \theta \)) is another solution of (2.2), then:

\[
\sum_{a=0}^{n} f_a(i) \cdot b[a; n, \theta] = b[i; s, \theta], i = 0, 1, 2, \ldots, s.
\]  

(2.3)
By subtracting (2.3) from (2.2) we get:

\[
\sum_{a=0}^{n} [p_a(i) - f_a(i)] \binom{n}{a} \theta^a (1 - \theta)^{n-a} = E[u_a(i)] = 0
\]

\[i = 0, 1, 2, \ldots, s\], where \(u_a(i) = p_a(i) - f_a(i)\).

Since \(u_a(i)\) is independent of \(\theta\) and it is well known that the family of binomial distributions \(B(n, \theta)\) is complete, then \(u_a(i) = 0\). Hence, \(p_a(i) = f_a(i)\) for \(i = 0, 1, 2, \ldots, s\).

2.1. Remarks

If in Theorem 2.1 we consider \(c = -1\), then we obtain a characterization theorem of the hypergeometric distribution. In this case, if we consider \(c' = 0\), the characterization theorem given by Skibinsky [19] is obtained as a particular case.

3. Characterization based on Pearson difference equation

Lemma 3.1. Consider the family of Polya distributions \(P(N, A, s, c)\), when \(c < 0\), and support \(T = [m, M]\), \(0 \leq m < M\), and let \(\{p_i\}\) be its pmf. Then,

\[p_i = \frac{(m - M)_{i-m} \cdot (m + M + \left(\frac{A}{c} - 1\right) + \left(\frac{-B}{c} - s + 1\right))_{i-m}}{(i-m)! \cdot (\left(\frac{-B}{c} - s + 1\right) + 2m)_{i-m}} \cdot p_m\]

where \(B = N - A\).

Proof. In Ollero and Ramos [11] (Cor. 2.3), it is proved that the probability generating function, \(G(z)\), of the Polya distribution \(P(N, A, s, c)\), is

\[G(z) = p_m \cdot z^m \cdot {}_2F_1\left(m - s, m + \frac{A}{c}; \frac{-B}{c} - s + 1 + 2m; z\right),\]

or, equivalently

\[G(z) = p_m \cdot z^m \cdot \sum_{i=0}^{M-m} \frac{(m - s)_i \cdot (m + A)_i}{\left(\frac{-B}{c} - s + 1 + 2m\right)_i} \cdot z^i \cdot i!\]

where \(_2F_1\) is the Gaussian hypergeometric function. In consequence,

\[p_{m+i} = p_m \cdot \frac{(m - s)_i \cdot (m + A)_i}{\left(\frac{-B}{c} - s + 1 + 2m\right)_i \cdot i!}\]

\[i = 0, 1, \ldots, M - m\].

The above relationship is equivalent to

\[p_i = p_m \cdot \frac{(m - s)_{i-m} \cdot (m + A)_{i-m}}{\left(\frac{-B}{c} - s + 1 + 2m\right)_{i-m} \cdot (i-m)!}\]
\[ p_i = \frac{(m - M)_{i-m} \cdot (m + M + \left(\frac{A}{c} - 1\right) + \left(\frac{B}{c} - s + 1\right))_{i-m}}{(i-m)! \cdot \left(\frac{B}{c} - s + 1\right) + 2m}_{i-m} \cdot p_m \]  

(3.1)

where \( M = \min\left(s, \frac{-A}{c}\right) \).

**Lemma 3.2.** If \( \alpha \) and \( \beta \) are the roots of the equation

\[ b_2 x^2 - (b_1 + b_2 - 1)x + b_1 + a - 1 = 0, \]  

(3.2)

then \(-\alpha + k + 1\) and \(-\beta + k + 1\) are the roots of

\[ b_2(x - k)^2 + (b_1 - b_2 - 1)(x - k) + a = 0. \]  

(3.3)

**Proof.** Since \( \alpha \) and \( \beta \) are the roots of (3.2), then

\[ b_2 \alpha^2 - (b_1 + b_2 - 1)\alpha + a + b_1 - 1 = 0 \]

and

\[ b_2 \beta^2 - (b_1 + b_2 - 1)\beta + a + b_1 - 1 = 0. \]

By substituting \(-\alpha + k + 1\) and \(-\beta + k + 1\) in (3.3) the result is easily proved. \( \square \)

Now, consider the subclass \( \mathcal{P}_H \) of the discrete Pearson system defined in Ollero and Ramos [11].

**Lemma 3.3.** If \( F \in \mathcal{P}_H \), then

\[ (m + 1)_{i-m} \left( m + \frac{b_1}{b_2} \right)_{i-m} = (i-m)! \left( \frac{b_1}{b_2} + 2m \right)_{i-m}. \]

**Proof.** Ollero and Ramos [11] proved that if \( F \in \mathcal{P}_H \), then

\[ m = 0 \text{ or } m = 1 - \frac{b_1}{b_2}, \]

and in both cases Lemma 3.3 is obtained:

\( \text{(a)} \) \( m = 0 \Rightarrow (m + 1)_{i-m} \left( m + \frac{b_1}{b_2} \right)_{i-m} = (i) \left( \frac{b_1}{b_2} \right)_{i-m} = (i-m)! \left( \frac{b_1}{b_2} + 2m \right)_{i-m} \)

\( \text{(b)} \) \( m = 1 - \frac{b_1}{b_2} \Rightarrow (m + 1)_{i-m} \left( m + \frac{b_1}{b_2} \right)_{i-m} = \left( \frac{b_1}{b_2} \right)_{i-m} + (i-m)! \left( \frac{b_1}{b_2} + 2m \right)_{i-m} \)

\( \square \)
Theorem 3.1. Let $F$ be a distribution and \{\(p_i\)\} its pmf. Then $F$ is the Pólya distribution $P(N,A,s,c)$, $c < 0$, if and only if

(i) $\Delta p_{i-1} = \frac{a-i}{b_1 i + b_2 i(i-1)} \cdot p_{i-1}$;
(ii) the support of $F$ is $T = \{m, m+1, \ldots, M\}$, $m \in \mathbb{N}$;
(iii) $b_2 \neq 0$;
(iv) $\frac{b_1}{b_2} + 2m \in \mathbb{R}^+$.

Proof.

(a) If $F \sim P(N,A,s,c)$, $c < 0$, from Ollero and Ramos [11] we conclude that $F \in \mathcal{P}_H$ and conditions (i) to (iv) are, in consequence, verified.

(b) Let $F$ be a distribution and \{\(p_i\)\} its pmf. Consider that conditions (i) to (iv) are verified, then $F \in \mathcal{P}_H$. From (i) we have that

$$p_i = \frac{a-i}{b_1 i + b_2 i(i-1)} \cdot p_{i-1} + p_i = \frac{b_2 i^2 + i(b_1 - b_2 - 1) + a}{b_2 i^2 + i(b_1 - b_2)} \cdot p_{i-1}
= \frac{b_2 i^2 + i(b_1 - b_2 - 1) + a}{b_2 i^2 + i(b_1 - b_2)} \cdot \frac{b_2(i-1)^2 + (i-1)(b_1 - b_2 - 1) + a}{b_2(i-1)^2 + (i-1)(b_1 - b_2)} \cdot p_{i-2}
= \frac{b_2 i^2 + i(b_1 - b_2 - 1) + a}{b_2 i^2 + i(b_1 - b_2)} \cdot \frac{b_2(m+1)^2 + (m+1)(b_1 - b_2 - 1) + a}{b_2(m+1)^2 + (m+1)(b_1 - b_2)} \cdot p_m
$$

for all $i = m, m+1, m+2, \ldots, M$. If $\alpha$ and $\beta$ are the roots of (3.2), from Lemma 3.2, $p_i$ can also be written as

$$p_i = \frac{b_2^{i-m}(i+\alpha - 1)\cdots(m+\alpha)(i+\beta - 1)\cdots(m+\beta)}{b_2^m i(i-1)\cdots(m+1)(i+\frac{b_1}{b_2} - 1)\cdots(m+\frac{b_1}{b_2})} \cdot p_m.$$

Hence,

$$p_i = \frac{(m+\alpha)_{i-m}(m+\beta)_{i-m}}{(m+1)_{i-m}(m+\frac{b_1}{b_2})_{i-m}} \cdot p_m
$$

$i = m, m+1, m+2, \ldots, M$. From Lemma 3.3,

$$p_i = \frac{(m+\alpha)_{i-m}(m+\beta)_{i-m}}{(i-m)!(\frac{b_1}{b_2} + 2m)_{i-m}} \cdot p_m.$$

Ollero and Ramos [11] (Cor. 2.2), proved that $\alpha = -M$ and $\beta = M + 1 - \frac{1 - b_1}{b_2}$, so:

$$p_i = \frac{(m-M)_{i-m}(m+M+1 - \frac{1}{b_2} + \frac{b_1}{b_2})}{(i-m)!(\frac{b_1}{b_2} + 2m)_{i-m}} \cdot p_m. \tag{3.4}$$

It is easy to show that (3.1) equal (3.4), with

$$\frac{b_1}{b_2} = \frac{-B}{c} - s + 1 \quad (\Leftrightarrow B = \left(1 - \frac{b_1}{b_2} - s\right) c)
1 - \frac{1}{b_2} = \frac{N}{c} - 1 \quad (\Leftrightarrow N = \left(2 - \frac{1}{b_2}\right) c).$$
Since \( A = N - B \), then
\[
F \sim P \left( \left( 2 - \frac{1}{b_2} \right) c, \left( 1 - \frac{1 - b_1}{b_2} + s \right) c, s, c \right).
\]

In order to determine \( s \), from (1.2) we obtain
\[
a = \left[ (1 - \frac{1 - b_1}{b_2} + s) c - c \right] (s + 1) \left( \frac{2}{c} - \frac{1}{c^2} \right) - 2c 
\]
\[
\Leftrightarrow a = \frac{(1 - \frac{1 - b_1}{b_2} + s) c (s + 1)}{c^2} \Leftrightarrow s^2 b_2 + s (b_1 + b_2 - 1) + b_1 + a - 1 = 0.
\]

Thus, \(-s\) is a root of \( b_2 x^2 - (b_1 + b_2 - 1) x + b_1 + a - 1 = 0 \).

Since the roots of \( b_2 x^2 - (b_1 + b_2 - 1) x + b_1 + a - 1 = 0 \) are \( \alpha = -M \) and \( \beta = M + 1 - \frac{1 - b_1}{b_2} \), then \( F \) has a Pólya distribution
\[
P_1 \equiv P \left( \left( 2 - \frac{1}{b_2} \right) c, \left( 1 - \frac{1 - b_1}{b_2} + M \right) c, M, c \right)
\]
\[
or
P_2 \equiv P \left( \left( 2 - \frac{1}{b_2} \right) c, -Mc, \frac{1 - b_1}{b_2} - M - 1, c \right),
\]

and from Note 1.2, \( P_1 \equiv d \ P_2 \).

In conclusion, the family \( P_H \) defined by Ollero and Ramos [11] as a subfamily of the discrete Pearson system, is the family of Pólya distributions with negative contagion parameter. \( \square \)

References