UCD GEARY INSTITUTE
DISCUSSION PAPER SERIES

Uniform Correlation Structure and Convex Stochastic Ordering in the Pólya urn Scheme

Raffaella Calabrese
Dynamics Lab, Geary Institute, University College Dublin

Geary WP2012/16
June 2012

UCD Geary Institute Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

Any opinions expressed here are those of the author(s) and not those of UCD Geary Institute. Research published in this series may include views on policy, but the institute itself takes no institutional policy positions.

Uniform correlation structure and convex stochastic ordering in the Pólya urn scheme

Raffaella Calabrese
raffaella.calabrese@ucd.ie
Dynamic Labs
Geary Institute
University College Dublin

Abstract

This analysis of the Pólya urn scheme focuses on proving two new important results, two new important results. First, the correlation structure between draws of a Pólya urn scheme is uniform. Second, fixing the number of draws and the composition of the urn, a convex stochastic ordering is induced by the number of balls replaced into the urn.

Keywords: Pólya urn, convex stochastic ordering, uniform correlation

1. Introduction

Urn models have been among the most popular probabilistic schemes and have been used to represent some form of contagion (see Johnson et al., 1997; Feller, 1968). The Pólya urn (Eggenberger and Pólya, 1923) has been originally applied to problems dealing with the spread of a contagious disease (see Johnson and Kotz, 1977; Marshall and Olkin, 1993). Later the model has been applied to a variety of different areas, examples of applications include modeling population growth (Blackwell and Kendall, 1964), sequential clinical trials, biology, industry and finance (Johnson and Kotz, 1977).

The aim of this paper is to prove two important results concerning the Pólya urn scheme. The first one is given by the uniform correlation structure between draws. The second one is the convex stochastic ordering induced by the number of balls replaced in the urn, fixing the number of draws and the composition of the urn.

This paper is organized as follows. The next section describes the key features of a Pólya urn scheme. Section 3 proves that the Pólya urn scheme is characterized by the uniform correlation structure. Finally, the subsequent section analyses the convex stochastic ordering in the Pólya urn scheme.

2. Pólya urn model

Urn models representing some form of contagion can be constructed in an unlimited variety of ways. The development of these methods stems largely from the Pólya-Eggenberger model (Eggenberger and Pólya, 1923). In this scheme an urn contains b black balls and r red balls. A ball is drawn at random and then replaced, together with s balls of the same color. The procedure is repeated n times.

Let X be the random variable that represents the number of drawn black balls. The number of the possible sequences of the results of n draws with k drawn black balls and $n-k$ red drawn balls is given by the binomial coefficient of n over k. For this reason and because each sequence presents the same probability, the following equation is obtained

$$P\{X=k\} = \left(\begin{array}{c} n \\ k \end{array} \right) \frac{b(b+s)...[b+(k-1)s]r(r+s)...[r+(n-k-1)s]}{(b+r)(b+r+s)...[b+r+(n-1)s]} \tag{2.1}$$

with $k = 0, 1, ..., n$.

Many equations express the probability distribution of the random variable X in a more compact way. By considering the ascendant factorial

$$x^{[r]} = x(x+1)...(x+r-1),$$

Johnson and Kotz (1977, pp. 178), for example, define $\alpha = b/s$ and $\beta = r/s$, obtaining

$$P\{X=k\} = \left(\begin{array}{c} n \\ k \end{array} \right) \frac{\alpha^{[k]}\beta^{[n-k]}}{(\alpha+\beta)^{[n]}}. \tag{2.2}$$

3. The uniform correlation structure

In the Pólya scheme the probability of drawing a black ball remains constant to $b/(b+r)$ from draw to draw (Johnson and Kotz, 1977). It follows

that the random variable X could be represented as the sum of n binary random variables identically distributed $X = \sum_{i=1}^{n} A_i$ with

$$A_i = \begin{cases} 1 & \dfrac{b}{b+r} \\ \\ 0 & \dfrac{r}{b+r} \end{cases} \qquad \forall \;\; i = 1, 2, ..., n. \tag{3.1}$$

Theorem 3.1. *In the Pólya urn model the linear correlation coefficient between every pair of draws remains constant to* $s/(b+r+s)$.

Proof. Since the n binary random variables (3.1) are identically distributed, the linear correlation coefficient between two draws is

$$r(A_i, A_u) = \frac{(b+r)^2 P\{(A_i = 1) \cap (A_u = 1)\} - b^2}{br} \tag{3.2}$$

for $i, u = 1, 2, ..., n$ and $i \neq u$.

To compute the probability $P\{(A_i = 1) \cap (A_u = 1)\}$ in the expression (3.2), the u-th draw is considered as the $(i+t)$-th draw with $1 \leq t \leq n-i$. By applying the theorem of total probability and by considering the probability function (2.2), the following equation is obtained

$$\begin{aligned} P\{(A_i = 1) \cap (A_u = 1)\} &= P\{(A_i = 1) \cap (A_{i+t} = 1)\} \\ &= \sum_{k=2}^{i+t} \left(\begin{array}{c} i+t-2 \\ k-2 \end{array} \right) \frac{\alpha^{[k]} \beta^{[i+t-k]}}{(\alpha+\beta)^{[i+t]}} \\ &= \frac{\alpha(\alpha+1)}{(\alpha+\beta)^{[i+t]}} (\alpha+\beta+2)^{[i+t-2]} \\ &= \frac{b}{b+r} \frac{b+s}{b+r+s}. \end{aligned} \tag{3.3}$$

In the equations (3.3), the change of variable $z = k-2$ is made and Newton's binomial series (Riordan, 1980, pp. 9) is applied.

By substituting the result (3.3) in the expression (3.2), the linear correlation coefficient is $r(A_i, A_u) = \dfrac{s}{b+r+s}$ for all $i, u = 1, 2, ..., n$, with $i \neq u$. □

4. Convex stochastic ordering

It is considered the following stochastic ordering between the cumulative distribution functions of random variables with the same expectation.

Definition 4.1 (Strictly convex ordering)**.** Let X and Y be two random variables such that[1]

$$E[\phi(X)] > E[\phi(Y)] \qquad \text{for all convex functions } \phi: \quad \mathbb{R} \to \mathbb{R} \tag{4.1}$$

provided the expectations exist. Then X is said be *strictly bigger than* Y *in the convex ordering* (denoted as $X \succ_{cx} Y$).

Convex functions are functions that take on their (relatively) larger values over regions of the form $(-\infty, a) \cup (b, \infty)$ for $a < b$. Therefore, if (4.1) holds, X is more likely to take "extreme" values than Y, it follows that X is "more variable" than Y.

The following theorem provides another characterization of the convex ordering.

Theorem 4.1. *Let X and Y be two random variables such that $E(X) = E(Y)$. Then $X \succ_{cx} Y$ if and only if*

$$\int_{-\infty}^{u} F_X(t)dt \geq \int_{-\infty}^{u} F_Y(t)dt \qquad \forall u \in \mathbb{R} \tag{4.2}$$

where the inequality is strictly satisfied for at least one point $x \in \mathbb{R}$.

Proof. See Shaked and Shanthikumar (2007), pp. 111. □

If the random variables X and Y represent the number of drawn black balls in the Pólya urn scheme, the following theorems show the application of the convex stochastic ordering to the Pólya urn scheme.

Theorem 4.2. *Let X and Y be two random variables that represent the number of times a black ball is drawn in n draws in two Pólya schemes with respectively s and s^* numbers of the balls that are replaced into the urn with $s > s^*$.*

[1] It is sufficient to consider only functions ϕ that are convex on the union of the supports of X and Y.

The ratio $\dfrac{P\{Y=k\}}{P\{X=k\}}$ *is a strictly monotone increasing function of* k

- *for* $0 \le k < n\dfrac{b}{b+r} + \dfrac{r}{b+r}$ *if* $s>0$ *and* $s^*>0$
- *for* $\max\left(0, n+\dfrac{r}{s^*}\right) \le k < n\dfrac{b}{b+r} + \dfrac{r}{b+r}$ *if* $s>0$ *and* $s^*<0$
  or if $s<0$ *and* $s^*<0$.

(4.3)

and it is a strictly monotone decreasing function of k

- *for* $n \ge k > n\dfrac{b}{b+r} + \dfrac{r}{b+r}$ *if* $s>0$ *and* $s^*>0$
- *for* $\min\left(n, -\dfrac{b}{s^*}\right) \ge k > n\dfrac{b}{b+r} + \dfrac{r}{b+r}$
  if $s>0$ *and* $s^*<0$ *or if* $s<0$ *and* $s^*<0$.

(4.4)

If $n\dfrac{b}{b+r} + \dfrac{r}{b+r}$ *is an integer number, it results*

$$\frac{P\{Y=k\}}{P\{X=k\}} = \frac{P\{Y=k-1\}}{P\{X=k-1\}} \qquad \text{for } k = n\frac{b}{b+r} + \frac{r}{b+r}.$$

Proof. Proving that the ratio $\dfrac{P\{Y=k\}}{P\{X=k\}}$ is a strictly monotone decreasing function of k for $k < n\dfrac{b}{b+r} + \dfrac{r}{b+r}$ is equivalent to proving that

$$\frac{P\{Y=k\}}{P\{X=k\}} > \frac{P\{Y=k-1\}}{P\{X=k-1\}} \quad \text{for } k < n\frac{b}{b+r} + \frac{r}{b+r}. \tag{4.5}$$

In turn, equation (4.5) is equivalent to the following condition

$$\frac{P\{Y=k\}}{P\{Y=k-1\}} > \frac{P\{X=k\}}{P\{X=k-1\}} \quad \text{for } k < n\frac{b}{b+r} + \frac{r}{b+r}.$$

From the expression (2.1) the following result is deduced

$$\frac{P\{X=k\}}{P\{X=k-1\}} = \frac{\binom{n}{k}}{\binom{n}{k-1}} \frac{b+(k-1)s}{r+(n-k)s}.$$

By applying the previous result, the following ratio is computed

$$\frac{P\{Y=k\}}{P\{Y=k-1\}}\frac{P\{X=k-1\}}{P\{X=k\}}=\frac{b+(k-1)s^*}{r+(n-k)s^*}\frac{r+(n-k)s}{b+(k-1)s}. \tag{4.6}$$

Knowing that $s > s^*$, in order that the probabilities in the equation (4.6) are non-null, the analysis is constrained to the following intervals

- $0 < k < n$ if $s > 0$ and $s^* > 0$;
- $\max\left(0, n+\frac{r}{s^*}\right) < k < \min\left(n, -\frac{b}{s^*}\right)$

  if $s > 0$ and $s^* < 0$ or if $s < 0$ and $s^* < 0$.

Determining for which values of k $P\{Y=k\}/P\{X=k\}$ is a strictly monotone increasing function of k is equivalent to determining for which values of k the ratio (4.6) is higher than one. This means that

$$\frac{b+(k-1)s^*}{r+(n-k)s^*} > \frac{b+(k-1)s}{r+(n-k)s}. \tag{4.7}$$

In the above-mentioned intervals the denominators of inequality (4.7) are always positive, hence

$$s[b(n-k)-r(k-1)] > s^*[b(n-k)-r(k-1)].$$

Since $s > s^*$, it is obtained

$$k < n\frac{b}{b+r}+\frac{r}{b+r}.$$

This means that $P\{Y=k\}/P\{X=k\}$ is a strictly monotone increasing function of k in the intervals (4.3).

Similarly, it is proven that $P\{Y=k\}/P\{X=k\}$ is a strictly monotone decreasing function of k in the intervals (4.4).

Finally, if $n\frac{b}{b+r}+\frac{r}{b+r}$ is an integer number, the ratio (4.6) is equal to one for $k = n\frac{b}{b+r}+\frac{r}{b+r}$. □

It is important to note that $n\frac{b}{b+r}$ is the expectation of both the random variables X and Y.

Lemma 4.3. [2] *Let X and Y be two discrete non-negative random variables with the same support $D = \{0, 1, ..., n\}$. If the following condition is satisfied*

$$F_X(k) \neq F_Y(k) \quad \textit{for } k \in I \subseteq D$$

and the expectations of X and Y coincide $E(X) = E(Y)$, then the function $P\{X = k\} - P\{Y = k\}$ have at least two changes of sign on the set D.

Proof. Knowing that

$$\sum_{k \in D} P\{X = k\} = \sum_{k \in D} P\{Y = k\} = 1,$$

and since $F_X(k) \neq F_Y(k)$ for $k \in I \subseteq D$, then a point k_0 exists in which the function $P\{X = k\} - P\{Y = k\}$ has a change of sign.

It is assumed that the function $P\{X = k\} - P\{Y = k\}$ has only one change of sign on the set D. Under this assumption, it results that

$$\begin{aligned} P\{X = k\} - P\{Y = k\} &\geq 0 \quad \text{for } \ k \leq k_0 \\ P\{X = k\} - P\{Y = k\} &\leq 0 \quad \text{for } \ k > k_0. \end{aligned}$$

This means that

$$(k - k_0)[P\{X = k\} - P\{Y = k\}] \leq 0 \qquad \forall k \in D$$

and at least one point $k \in D$ satisfies the following condition

$$(k - k_0)[P\{X = k\} - P\{Y = k\}] < 0. \tag{4.8}$$

From the inequality (4.8) it can be deduced that

$$\sum_{k=0}^{n} (k - k_0)[P\{X = k\} - P\{Y = k\}] = E(X) - E(Y) < 0.$$

This result is in contrast with the assumption that the expectations coincide $E(X) = E(Y)$.
Conversely, if the following condition is satisfied

$$\begin{aligned} P\{X = k\} - P\{Y = k\} &\leq 0 \quad \text{for } \ k \leq k_0 \\ P\{X = k\} - P\{Y = k\} &\geq 0 \quad \text{for } \ k > k_0, \end{aligned}$$

[2] For a generalization of this lemma see Denuit and Lefévre (1997).

the following will likewise be true

$$\sum_{k=0}^{n}(k-k_0)[P\{X=k\}-P\{Y=k\}]=E(X)-E(Y)>0.$$

Moreover, this result is in contrast with the assumption that the expectations $E(X)=E(Y)$ coincide.
Then, it can be deduced that the function $P\{X=k\}-P\{Y=k\}$ has at least two changes of sign. □

The previous result allows to prove the following theorem.

Theorem 4.4. *Let X and Y be two random variables that represent the number of times a black ball is drawn in n draws in two Pólya schemes with respectively s and s^* numbers of the balls that are replaced into the urn with $s > s^*$. The function $P\{X=k\}-P\{Y=k\}$ has two changes of sign and the sign sequence is $+$, $-$, $+$ in the following sets*

- $0 \le k \le n$ *if* $s>0$ *and* $s^*>0$ *or if* $s>0$ *and* $s^*<0$*;*
- $\max\left(0, n+\frac{r}{s}\right) \le k \le \min\left(n, -\frac{b}{s}\right)$ *if* $s<0$ *and* $s^*<0$.

Proof. First, the case in which the random variables' supports are different is considered. Since $s>s^*$, this condition is satisfied only if $s>0$ and $s^*<0$ or $s<0$ and $s^*<0$. In both the cases the support of the random variable Y is included or coincides with the support of the random variable X. This means that $P\{X=k\}-P\{Y=k\}>0$ in the following intervals

- for $0 \le k < \max\left(0, n+\frac{r}{s^*}\right)$ and $\min\left(n, -\frac{b}{s^*}\right) < k \le n$

  if $s>0$ and $s^*<0$;
- for $\max\left(0, n+\frac{r}{s}\right) \le k < \max\left(0, n+\frac{r}{s^*}\right)$ and

  $\min\left(n, -\frac{b}{s^*}\right) < k \le \min\left(n, -\frac{b}{s}\right)$

  if $s<0$ and $s^*<0$.

The supports of the random variables X and Y are denoted by, respectively, D_1 and D_2. Since

$$\sum_{k\in D_1} P\{X=k\} = \sum_{k\in D_2} P\{Y=k\} = 1,$$

the following condition $P\{X=k\} - P\{Y=k\} < 0$ must be satisfied for some values of $k \in D_1 \cap D_2$.
From the lemma (), only one interval I exists such that

$$\frac{P\{Y=k\}}{P\{X=k\}} > 1 \quad \text{for } k \in I \subseteq \left[\max\left(0, n+\frac{r}{s^*}\right), \min\left(n, -\frac{b}{s^*}\right)\right]. \tag{4.9}$$

The condition (4.9) coincides with the following condition

$$P\{X=k\} - P\{Y=k\} < 0 \quad \text{for } \ k \in I \subseteq \left[\max\left(0, n+\frac{r}{s^*}\right), \min\left(n, -\frac{b}{s^*}\right)\right].$$

This means that the function $P\{X=k\} - P\{Y=k\}$ has two changes of sign on the support D_1 with the sign sequence $+, -, +$.

Second, the case in which the random variables' supports coincide $D = \{0, 1, \ldots, n\}$, is considered. Since $\mathbb{E}(X) = \mathbb{E}(Y)$ and $s \neq s^*$, from the probability function (2.1) it can be deduced that $F_X(k) \neq F_Y(k)$ for $k \in B \subseteq D$. This means that the assumptions of lemma (4) are satisfied. From this lemma the function $P\{X=k\} - P\{Y=k\}$ has at least two changes of sign on the set D.

From the lemma (4.2), the function $P\{X=k\} - P\{Y=k\}$ has two changes of sign on the set D at the points $k_0,\ k_1 \in D$. Moreover, it results

$$\frac{P\{Y=k\}}{P\{X=k\}} \leq 1 \quad \text{for } k \leq k_0 \text{ and } k \geq k_1,$$

with $0 \leq k_0 < n\dfrac{b}{b+r} + \dfrac{r}{b+r}$ and $n \geq k_1 > n\dfrac{b}{b+r} + \dfrac{r}{b+r}$.

Since

$$\sum_{k=0}^{n} P\{X=k\} = \sum_{k=0}^{n} P\{Y=k\} = 1, \tag{4.10}$$

at least one point $k \in D$ exists such that

$$\frac{P\{Y=k\}}{P\{X=k\}} > 1 \quad \text{for } k_0 < k < k_1.$$

This means that

$$\begin{aligned} &P\{X=k\}-P\{Y=k\}\geq 0 \quad \text{for} \quad k\leq k_0 \text{ and } k\geq k_1 \\ &P\{X=k\}-P\{Y=k\}<0 \quad \text{for} \quad k_0<k<k_1. \end{aligned} \tag{4.11}$$

From the lemma (4.2) or from the condition (4.10) it results that the first inequality in (4.11) must be satisfied with strict inequality for at least one point $k \in D$. □

From the theorems (4.1) and (4.4) and by knowing that the random variables X and Y have the same expectations, it can be deduced that X is strictly bigger than Y in the convex ordering.

5. References

Blackwell D. and Kendall D. (1964) The Martin boundary for Pólyas urn scheme, and an application to stochastic population growth.*Journal of Applied Probability*, **1**, 284-296.

Denuit M. and Lefévre C. (1997). Some new classes of stochastic order relations among arithmetic random variables, with applications in actuarial sciences. *Insurance: Mathematics & Economics* **20**, 197-213.

Eggenberger F. and Polya G. (1923). Uber die Statistik verketter Vorgange. Zeitschrift fur Angewandte Mathematik und Mechanik, **3**, 279-289.

Feller W. (1968). *An Introduction to Probability Theory and Its Applications.* Vol. I, 3rd ed. John Wiley & Sons, New York.

Johnson N. L. and Kotz S. (1977). *Urn Models and Their Application.* John Wiley & Sons, New York.

Marshall A. W. and Olkin I. (1993). Bivariate life distributions from Pólya's urn model for contagion, Journal of Applied Probability, **30**, 497508.

Riordan J. (1980). *An Introduction to Combinatorial Analysis.* Princeton University Press, Princeton.

Shaked M. and Shanthikumar J. G. (1994). *Stochastic Orders and Their Applications.* Academic Press, San Diego.