# 12.8: Pólya's Urn Process

## Basic Theory

### The Model

Pólya's urn scheme is a dichotomous sampling model that generalizes the hypergeometric model (sampling without replacement) and the Bernoulli model (sampling with replacement). Pólya's urn proccess leads to a famous example of a sequence of random variables that is exchangeable, but not independent, and has deep conections with the beta-Bernoulli process.

Suppose that we have an urn (what else!) that initially contains $a$ red and $b$ green balls, where $a$ and $b$ are positive integers. At each discrete time (trial), we select a ball from the urn and then return the ball to the urn along with $c$ new balls of the same color. Ordinarily, the parameter $c$ is a nonnegative integer. However, the model actually makes sense if $c$ is a negative integer, if we interpret this to mean that we *remove* the balls rather than add them, and assuming that there are enough balls of the proper color in the urn to perform this action. In any case, the random process is known as *Pólya's urn process*, named for George Pólya.

In terms of the colors of the selected balls, Pólya's urn scheme generalizes the standard models of sampling with and without replacement.

1. $c = 0$ corresponds to sampling with replacement.
2. $c = -1$ corresponds to sampling without replacement.

For the most part, we will assume that $c$ is nonnegative so that the process can be continued indefinitely. Occasionally we consider the case $c = -1$ so that we can interpret the results in terms of sampling without replacement.

### The Outcome Variables

Let $X_i$ denote the color of the ball selected at time $i$, where 0 denotes green and 1 denotes red. Mathematically, our basic random process is the sequence of indicator variables $\boldsymbol{X} = (X_1, X_2, \ldots)$, known as the *Pólya process*. As with any random process, our first goal is to compute the *finite dimensional distributions* of $\boldsymbol{X}$. That is, we want to compute the joint distribution of $(X_1, X_2, \ldots, X_n)$ for each $n \in \mathbb{N}_+$. Some additional notation will really help. Recall the generalized permutation formula in our study of combinatorial structures: for $r,\ s \in \mathbb{R}$ and $j \in \mathbb{N}$, we defined

$$r^{(s,j)} = r(r+s)(r+2s)\cdots[r+(j-1)s] \tag{12.8.1}$$

Note that the expression has $j$ factors, starting with $r$, and with each factor obtained by adding $s$ to the previous factor. As usual, we adopt the convention that a product over an empty index set is 1. Hence $r^{(s,0)} = 1$ for every $r$ and $s$.

Recall that

1. $r^{(0,j)} = r^j$, an ordinary power
2. $r^{(-1,j)} = r^{(j)} = r(r-1)\cdots(r-j+1)$ , a descending power
3. $r^{(1,j)} = r^{[j]} = r(r+1)\cdots(r+j-1)$ , an ascending power
4. $r^{(r,j)} = j! r^j$
5. $1^{(1,j)} = j!$

The following simple result will turn out to be quite useful.

Suppose that $r,\ s \in (0,\infty)$ and $j \in \mathbb{N}$. Then

$$\frac{r^{(s,j)}}{s^j} = \left(\frac{r}{s}\right)^{[j]} \tag{12.8.2}$$

Proof

It's just a matter of grouping the factors:

$$\begin{aligned}\frac{r^{(s,j)}}{s^j} &= \frac{r(r+s)(r+2s)\cdots[r+(j-1)s]}{s^j} \\ &= \left(\frac{r}{s}\right)\left(\frac{r}{s}+1\right)\left(\frac{r}{s}+2\right)\cdots\left[\frac{r}{s}+(j-1)\right] = \left(\frac{r}{s}\right)^{[j]}\end{aligned}$$

The finite dimensional distributions are easy to compute using the multiplication rule of conditional probability. If we know the contents of the urn at any given time, then the probability of an outcome at the next time is all but trivial.

Let $n \in \mathbb{N}_+$, $(x_1, x_2, \ldots, x_n) \in \{0,1\}^n$ and let $k = x_1 + x_2 + \cdots + x_n$ . Then

$$\mathbb{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \frac{a^{(c,k)} b^{(c,n-k)}}{(a+b)^{(c,n)}} \tag{12.8.3}$$

Proof

By the multiplication rule for conditional probability,

$$\mathbb{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \mathbb{P}(X_1 = x_1)\mathbb{P}(X_2 = x_2 \mid X_1 = x_1)\cdots\mathbb{P}(X_n = x_n \mid X_1 = x_1, \ldots, X_{n-1} = x_{n-1}) \tag{12.8.4}$$

Of course, if we know that the urn has, say, $r$ red and $g$ green balls at a particular time, then the probability of a red ball on the next draw is $r/(r+g)$ while the probability of a green ball is $g/(r+g)$. The right side of the displayed equation above has $n$ factors. The denominators are the total number of balls at the $n$ times, and form the product $(a+b)(a+b+c)\ldots[a+b+(n-1)c] = (a+b)^{(c,n)}$. In the numerators, $k$ of the factors correspond to probabilities of selecting red balls; these factors form the product $a(a+c)\cdots[a+(k-1)c] = a^{(c,k)}$. The remaining $n-k$ factors in the numerators correspond to selecting green balls; these factors form the product $b(b+c)\cdots[b+(n-k-1)c] = b^{(c,n-k)}$.

The joint probability in the previous exercise depends on $(x_1, x_2, \ldots, x_n)$ only through the number of red balls $k = \sum_{i=1}^n x_i$ in the sample. Thus, the joint distribution is invariant under a permutation of $(x_1, x_2, \ldots, x_n)$, and hence $\boldsymbol{X}$ is an exchangeable sequence of random variables. This means that for each $n$, all permutations of $(X_1, X_2, \ldots, X_n)$ have the same distribution. Of course the joint distribution reduces to the formulas we have obtained earlier in the special cases of sampling with replacement ($c = 0$) or sampling without replacement ($c = -1$), although in the latter case we must have $n \le a+b$. When $c > 0$, the Pólya process is a special case of the beta-Bernoulli process, studied in the chapter on Bernoulli trials.

The Pólya process $\boldsymbol{X} = (X_1, X_2, \ldots)$ with parameters $a, b, c \in \mathbb{N}_+$ is the beta-Bernoulli process with parameters $a/c$ and $b/c$. That is, for $n \in \mathbb{N}_+$, $(x_1, x_2, \ldots, x_n) \in \{0,1\}^n$, and with $k = x_1 + x_2 + \cdots + x_n$,

$$\mathbb{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \frac{(a/c)^{[k]}(b/c)^{[n-k]}}{(a/c+b/c)^{[n]}} \tag{12.8.5}$$

Proof

From the previous two results,

$$\mathbb{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \frac{a^{(c,k)}b^{(c,n-k)}}{(a+b)^{(c,n)}} = \frac{[a^{(c,k)}/c^k][b^{(c,n-k)}/c^{n-k}]}{(a+b)^{(c,n)}/c^n} = \frac{(a/c)^{[k]}(b/c)^{[n-k]}}{(a/c+b/c)^{[n]}} \tag{12.8.6}$$

and this is the corresponding finite dimensional distribution of the beta-Bernoulli distribution with parameters $a/c$ and $(b/c)$.

Recall that the beta-Bernoulli process is obtained, in the usual formulation, by randomizing the success parameter in a Bernoulli trials sequence, giving the success parameter a beta distribution. So specifically, suppose $a, b, c \in \mathbb{N}_+$ and that random variable $P$ has the beta distribution with parameters $a/c$ and $b/c$. Suppose also that given $P = p \in (0,1)$, the random process $\boldsymbol{X} = (X_1, X_2, \ldots)$ is a sequence of Bernoulli trials with success parameter $p$. Then $\boldsymbol{X}$ is the Pólya process with parameters $a, b, c$. This is a fascinating connection between two processes that at first, seem to have little in common. In fact however, *every* exchangeable sequence of indicator random variables can be obtained by randomizing the success parameter in a sequence of Bernoulli trials. This is *de Finetti's theorem*, named for Bruno de Finetti, which is studied in the section on backwards martingales. When $c \in \mathbb{N}_+$, all of the results in this section are special cases of the corresponding results for the beta-Bernoulli process, but it's still interesting to interpret the results in terms of the urn model.

For each $i \in \mathbb{N}_+$

1. $\mathbb{E}(X_i) = \frac{a}{a+b}$
2. $\operatorname{var}(X_i) = \frac{a}{a+b}\frac{b}{a+b}$

Proof

Since the sequence is exchangeable, $X_i$ has the same distribution as $X_1$, so $\mathbb{P}(X_i = 1) = \frac{a}{a+b}$. The mean and variance now follow from standard results for indicator variables.

Thus $\boldsymbol{X}$ is a sequence of identically distributed variables, quite surprising at first but of course inevitable for any exchangeable sequence. Compare the joint and marginal distributions. Note that $\boldsymbol{X}$ is an independent sequence if and only if $c = 0$, when we have simple sampling with replacement. Pólya's urn is one of the most famous examples of a random process in which the outcome variables are exchangeable, but dependent (in general).

Next, let's compute the covariance and correlation of a pair of outcome variables.

Suppose that $i, j \in \mathbb{N}_+$ are distinct. Then

1. $\operatorname{cov}(X_i, X_j) = \frac{abc}{(a+b)^2(a+b+c)}$
2. $\operatorname{cor}(X_i, X_j) = \frac{c}{a+b+c}$

Proof

Since the variables are exchangeable, $\mathbb{P}(X_i = 1, X_j = 1) = \mathbb{P}(X_1 = 1, X_2 = 1) = \frac{a}{a+b}\frac{a+c}{a+b+c}$. The results now follow from standard formulas for covariance and correlation.

Thus, the variables are positively correlated if $c > 0$, negatively correlated if $c < 0$, and uncorrelated (in fact, independent), if $c = 0$. These results certainly make sense when we recall the dynamics of Pólya's urn. It turns out that in any *infinite* sequence of exchangeable variables, the variables must be nonnegatively correlated. Here is another result that explores how the variables are related.

Suppose that $n \in \mathbb{N}_+$ and $(x_1, x_2, \ldots, x_n) \in \{0,1\}^n$. Let $k = \sum_{i=1}^n x_i$. Then

$$\mathbb{P}(X_{n+1} = 1 \mid X_1 = x_1, X_2 = x_2, \ldots X_n = x_n) = \frac{a + kc}{a + b + nc} \tag{12.8.7}$$

Proof

Using the joint distribution,

$$\begin{aligned}\mathbb{P}(X_{n+1} = 1 \mid X_1 = x_1, X_2 = x_2, \ldots X_n = x_n) &= \frac{\mathbb{P}(X_1 = x_1, X_2 = x_2, \ldots X_n = x_n, X_{n+1} = 1)}{\mathbb{P}(X_1 = x_1, X_2 = x_2, \ldots X_n = x_n)} \\ &= \frac{a^{(c,k+1)} b^{(c,n-k)}}{(a+b)^{(c,n+1)}} \frac{(a+b)^{(c,n)}}{a^{(c,k)} b^{(c,n-k)}} = \frac{a + ck}{a + b + cn}\end{aligned}$$

Pólya's urn is described by a sequence of indicator variables. We can study the same derived random processes that we studied with Bernoulli trials: the number of red balls in the first $n$ trials, the trial number of the $k$th red ball, and so forth.

## The Number of Red Balls

For $n \in \mathbb{N}$, the number of red balls selected in the first $n$ trials is

$$Y_n = \sum_{i=1}^n X_i \tag{12.8.8}$$

so that $\boldsymbol{Y} = (Y_0, Y_1, \ldots)$ is the partial sum process associated with $\boldsymbol{X} = (X_1, X_2, \ldots)$.

Note that

1. The number of green balls selected in the first $n$ trials is $n - Y_n$.
2. The number of red balls in the urn after the first $n$ trials is $a + c\,Y_n$.
3. The number of green balls in the urn after the first $n$ trials is $b + c(n - Y_n)$.
4. The number of balls in the urn after the first $n$ trials is $a + b + c\,n$.

The basic analysis of $\boldsymbol{Y}$ follows easily from our work with $\boldsymbol{X}$.

The probability density function of $Y_n$ is given by

$$\mathbb{P}(Y_n = k) = \binom{n}{k} \frac{a^{(c,k)} b^{(c,n-k)}}{(a+b)^{(c,n)}}, \quad k \in \{0, 1, \ldots, n\} \tag{12.8.9}$$

Proof

$\mathbb{P}(Y_n = y)$ is the sum of $\mathbb{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$ over all $(x_1, x_2, \ldots, x_n) \in \{0,1\}^n$ with $\sum_{i=1}^n x_i = k$. There are $\binom{n}{k}$ such sequences, and each has the probability given above.

The distribution defined by this probability density function is known, appropriately enough, as the *Pólya distribution* with parameters $n$, $a$, $b$, and $c$. Of course, the distribution reduces to the binomial distribution with parameters $n$ and $a/(a+b)$ in the case of sampling with replacement ($c = 0$) and to the hypergeometric distribution with parameters $n$, $a$, and $b$ in the case of sampling without replacement ($c = -1$), although again in this case we need $n \le a + b$. When $c > 0$, the Póyla distribution is a special case of the beta-binomial distribution.

If $a,\ b,\ c \in \mathbb{N}_+$ then the Pólya distribution with parameters $a,\ b,\ c$ is the beta-binomial distribution with parameters $a/c$ and $b/c$. That is,

$$P(Y_n = k) = \binom{n}{k} \frac{(a/c)^{[k]} (b/c)^{[n-k]}}{(a/c + b/c)^{[n]}}, \quad k \in \{0, 1, \ldots, n\} \tag{12.8.10}$$

Proof

This follows immediately from the result above that $\boldsymbol{X} = (X_1, X_2, \ldots)$ is the beta-Bernoulli process with parameters $a/c$ and $b/c$. So by definition, $Y_n = \sum_{i=1}^n X_i$ has the beta-binomial distribution with parameters $n$, $a/c$, and $b/c$. A direct proof is also simple using the permutation formula above:

$$\mathbb{P}(Y_n = k) = \binom{n}{k} \frac{a^{(c,k)} b^{(c,n-k)}}{(a+b)^{(c,n)}} = \binom{n}{k} \frac{[a^{(c,k)}/c^k][b^{(c,n-k)}/c^{n-k}]}{(a+b)^{(c,n)}/c^n} = \binom{n}{k} \frac{(a/c)^{[k]} (b/c)^{[n-k]}}{(a/c + b/c)^{[n]}}, \quad k \in \{0, 1, \ldots, n\} \tag{12.8.11}$$

The case where all three parameters are equal is particularly interesting.

If $a = b = c$ then $Y_n$ is uniformly distributed on $\{0, 1, \ldots, n\}$.

Proof

This follows from the previous result, since the beta-binomial distribution with parameters $n$, 1, and 1 reduces to the uniform distribution. Specifically, note that $1^{[k]} = k!$, $1^{[n-k]} = (n-k)!$ and $2^{[n]} = (n+1)!$. So substituting gives

$$\mathbb{P}(Y_n = k) = \frac{n!}{k!(n-k)!} \frac{k!(n-k)!}{(n+1)!} = \frac{1}{n+1}, \quad k \in \{0, 1, \ldots, n\} \tag{12.8.12}$$

In general, the Pólya family of distributions has a diverse collection of shapes.

Start the simulation of the Pólya Urn Experiment. Vary the parameters and note the shape of the probability density function. In particular, note when the function is skewed, when the function is symmetric, when the function is unimodal, when the function is monotone, and when the function is U-shaped. For various values of the parameters, run the simulation 1000 times and compare the empirical density function to the probability density function.

The Pólya probability density function is

1. unimodal if $a > b > c$ and $n > \frac{a-c}{b-c}$
2. unimodal if $b > a > c$ and $n > \frac{b-c}{a-c}$
3. U-shaped if $c > a > b$ and $n > \frac{c-b}{c-a}$
4. U-shaped if $c > b > a$ and $n > \frac{c-a}{c-b}$
5. increasing if $b < c < a$
6. decreasing if $a < c < b$

Proof

These results follow from solving the inequality $\mathbb{P}(Y_n = k) > \mathbb{P}(Y_n = k-1)$.

Next, let's find the mean and variance. Curiously, the mean does not depend on the parameter $c$.

The mean and variance of the number of red balls selected are

1. $\mathbb{E}(Y_n) = n\frac{a}{a+b}$
2. $\text{var}(Y_n) = n\frac{ab}{(a+b)^2}\left[1 + (n-1)\frac{c}{a+b+c}\right]$

Proof

These results follow from the mean and covariance of the indicator variables given above, and basic properties of expected value and variance.

1. $\mathbb{E}(Y_n) = \sum_{i=1}^n \mathbb{E}(X_i)$
2. $\text{var}(Y_n) = \sum_{i=1}^n \sum_{j=1}^n \text{cov}(X_i, X_j)$

Start the simulation of the Pólya Urn Experiment. Vary the parameters and note the shape and location of the mean $\pm$ standard deviation bar. For various values of the parameters, run the simulation 1000 times and compare the empirical mean and standard deviation to the distribution mean and standard deviation.

Explicitly compute the probability density function, mean, and variance of $Y_5$ when $a = 6$, $b = 4$, and for the values of $c \in \{-1, 0, 1, 2, 3, 10\}$ Sketch the graph of the density function in each case.

Fix $a$, $b$, and $n$, and let $c \to \infty$. Then

1. $\mathbb{P}(Y_n = 0) \to \frac{b}{a+b}$
2. $\mathbb{P}(Y_n = n) \to \frac{a}{a+b}$
3. $\mathbb{P}(Y_n \in \{1, 2, \ldots, n-1\}) \to 0$

Proof

Note that $\mathbb{P}(Y_n = 0) = \frac{b^{(c,n)}}{(a+b)^{(c,n)}}$. The numerator and denominator each have $n$ factors. If these factors are grouped into a product of $n$ fractions, then the first is $\frac{b}{a+b}$. The rest have the form $\frac{a+j\,c}{a+b+j\,c}$ where $j \in \{1, 2, \ldots n-1\}$ Each of these converges to 1 as $c \to \infty$. Part (b) follows by a similar argument. Part (c) follows from (a) and (b) and the complement rule.

Thus, the limiting distribution of $Y_n$ as $c \to \infty$ is concentrated on 0 and $n$. The limiting probabilities are just the initial proportion of green and red balls, respectively. Interpret this result in terms of the dynamics of Pólya's urn scheme.

Our next result gives the conditional distribution of $X_{n+1}$ given $Y_n$.

Suppose that $n \in \mathbb{N}$ and $k \in \{0, 1, \ldots, n\}$. Then

$$\mathbb{P}(X_{n+1} = 1 \mid Y_n = k) = \frac{a + ck}{a + b + cn} \tag{12.8.13}$$

Proof

Let $S = \{(x_1, x_2, \ldots, x_n) \in \{0,1\}^n : \sum_{i=1}^n x_i = k\}$ and let $\boldsymbol{X}_n = (X_1, X_2, \ldots, X_n)$. Note that the events $\{\boldsymbol{X}_n = \boldsymbol{x}\}$ over $\boldsymbol{x} \in S$ partition the event $\{Y_n = k\}$. Conditioning on $\boldsymbol{X}_n$,

$$\begin{aligned}\mathbb{P}(X_{n+1} = 1 \mid Y_n = k) &= \sum_{\boldsymbol{x} \in S} \mathbb{P}(X_{n+1} = 1 \mid Y_n = k, \boldsymbol{X}_n = \boldsymbol{x}) \mathbb{P}(\boldsymbol{X}_n = \boldsymbol{x} \mid Y_n = k) \\ &= \sum_{\boldsymbol{x} \in S} \mathbb{P}(X_{n+1} = 1 \mid \boldsymbol{X}_n = \boldsymbol{x}) \mathbb{P}(\boldsymbol{X}_n = \boldsymbol{x} \mid Y_n = k)\end{aligned}$$

But from our result above, $\mathbb{P}(X_{n+1} = 1 \mid \boldsymbol{X}_n = \boldsymbol{x}) = (a + ck)/(a + b + cn) \quad$ for every $\boldsymbol{x} \in S$. Hence

$$\mathbb{P}(X_{n+1} = 1 \mid Y_n = k) = \frac{a + ck}{a + b + cn} \sum_{\boldsymbol{x} \in S} \mathbb{P}(\boldsymbol{X}_n = x \mid Y_n = k) \tag{12.8.14}$$

The last sum is 1.

In particular, if $a = b = c$ then $\mathbb{P}(X_{n+1} = 1 \mid Y_n = n) = \frac{n+1}{n+2}$ . This is *Laplace's rule of succession*, another interesting connection. The rule is named for Pierre Simon Laplace, and is studied from a different point of view in the section on Independence.

## The Proportion of Red Balls

Suppose that $c \in \mathbb{N}$, so that the process continues indefinitely. For $n \in \mathbb{N}_+$ , the proportion of red balls selected in the first $n$ trials is

$$M_n = \frac{Y_n}{n} = \frac{1}{n} \sum_{i=1}^n X_i \tag{12.8.15}$$

This is an interesting variable, since a little reflection suggests that it may have a limit as $n \to \infty$. Indeed, if $c = 0$, then $M_n$ is just the sample mean corresponding to $n$ Bernoulli trials. Thus, by the law of large numbers, $M_n$ converges to the success parameter $\frac{a}{a+b}$ as $n \to \infty$ with probability 1. On the other hand, the proportion of red balls in the urn after $n$ trials is

$$Z_n = \frac{a + cY_n}{a + b + cn} \tag{12.8.16}$$

When $c = 0$, of course, $Z_n = \frac{a}{a+b}$ so that in this case, $Z_n$ and $M_n$ have the same limiting behavior. Note that

$$Z_n = \frac{a}{a + b + cn} + \frac{cn}{a + b + cn} M_n \tag{12.8.17}$$

Since the constant term converges to 0 as $n \to \infty$ and the coefficient of $M_n$ converges to 1 as $n \to \infty$, it follows that the limits of $M_n$ and $Z_n$ as $n \to \infty$ will be the same, *if* the limit exists, for any mode of convergence: with probability 1, in mean, or in distribution. Here is the general result when $c > 0$.

Suppose that $a, b, c \in \mathbb{N}_+$. There exists a random variable $P$ having the beta distribution with parameters $a/c$ and $b/c$ such that $M_n \to P$ and $Z_n \to P$ as $n \to \infty$ with probability 1 and in mean square, and hence also in distribution.

Proof

As noted earlier, the urn process is equivalent to the beta-Bernoulli process with parameters $a/c$ and $b/c$. We showed in that section that $M_n \to P$ as $n \to \infty$ with probability 1 and in mean square, where $P$ is the beta random variable used in the construction.

In turns out that the random process $\boldsymbol{Z} = \{Z_n = (a + cY_n)/(a + b + cn) : n \in \mathbb{N}\}$ is a martingale. The theory of martingales provides powerful tools for studying convergence in Pólya's urn process. As an interesting special case, note that if $a = b = c$ then the limiting distribution is the uniform distribution on $(0, 1)$.

## The Trial Number of the $k$th Red Ball

Suppose again that $c \in \mathbb{N}$, so that the process continues indefinitely. For $k \in \mathbb{N}_+$ let $V_k$ denote the trial number of the $k$th red ball selected. Thus

$$V_k = \min\{n \in \mathbb{N}_+ : Y_n = k\} \tag{12.8.18}$$

Note that $V_k$ takes values in $\{k, k+1, \ldots\}$. The random processes $\boldsymbol{V} = (V_1, V_2, \ldots)$ and $\boldsymbol{Y} = (Y_1, Y_2, \ldots)$ are inverses of each other in a sense.

For $k, n \in \mathbb{N}_+$ with $k \le n$,

1. $V_k \le n$ if and only if $Y_n \ge k$
2. $V_k = n$ if and only if $Y_{n-1} = k - 1$ and $X_n = 1$

The probability denisty function of $V_k$ is given by

$$\mathbb{P}(V_k = n) = \binom{n-1}{k-1} \frac{a^{(c,k)} b^{(c,n-k)}}{(a+b)^{(c,n)}}, \quad n \in \{k, k+1, \ldots\} \tag{12.8.19}$$

Proof

We condition on $Y_{n-1}$. Using the PDF of $Y_{n-1}$ and the result above,

$$\begin{aligned}\mathbb{P}(V_k = n) &= \mathbb{P}(Y_{n-1} = k-1, X_n = 1) = \mathbb{P}(Y_{n-1} = k-1)\mathbb{P}(X_n = 1 \mid Y_{n-1} = k-1) \\ &= \binom{n-1}{k-1} \frac{a^{(c,k-1)} b^{(c,(n-1)-(k-1))}}{(a+b)^{(c,n-1)}} \frac{a + c(k-1)}{a+b+c(n-1)} = \binom{n-1}{k-1} \frac{a^{(c,k)} b^{(c,n-k)}}{(a+b)^{(c,n)}}\end{aligned}$$

Of course this probability density function reduces to the negative binomial density function with trial parameter $k$ and success parameter $p = \frac{a}{a+b}$ when $c = 0$ (sampling with replacement). When $c > 0$, the distribution is a special case of the beta-negative binomial distribution.

If $a, b, c \in \mathbb{N}_+$ then $V_k$ has the beta-negative binomial distribution with parameters $k$, $a/c$, and $b/c$. That is,

$$\mathbb{P}(V_k = n) = \binom{n-1}{k-1} \frac{(a/c)^{[k]} (b/c)^{[n-k]}}{(a/c + b/c)^{[n]}}, \quad n \in \{k, k+1, \ldots\} \tag{12.8.20}$$

Proof

As with previous proofs, this result follows since the underlying process $\boldsymbol{X} = (X_1, X_2, \ldots)$ is the beta-Bernoulli process with parameters $a/c$ and $b/c$. The form of the PDF also follows easily from the previous result by dividing the numerator and denominator $c^n$.

If $a = b = c$ then

$$\mathbb{P}(V_k = n) = \frac{k}{n(n+1)}, \quad n \in \{k, k+1, k+2, \ldots\} \tag{12.8.21}$$

Proof

As in the corresponding proof for the number of red balls, the fraction in the PDF of $V_k$ in the previous result reduces to $\frac{k!(n-k)!}{(n+1)!}$, while the binomial coefficient is $\frac{(n-1)!}{(k-1)!(n-k)!}$.

Fix $a$, $b$, and $k$, and let $c \to \infty$. Then

1. $\mathbb{P}(V_k = k) \to \frac{a}{a+b}$
2. $\mathbb{P}(V_k \in \{k+1, k+2, \ldots\}) \to 0$

Thus, the limiting distribution of $V_k$ is concentrated on $k$ and $\infty$. The limiting probabilities at these two points are just the initial proportion of red and green balls, respectively. Interpret this result in terms of the dynamics of Pólya's urn scheme.