JIRSS (2004)
Vol. 3, No. 2, pp 165-173

Pólya-Type Urn Models with Multiple Drawings

Norman Johnson[1], Samuel Kotz[2], Hosam Mahmoud[2]

[1]University of North Carolina, Chapel Hill, North Carolina, U.S.A.
[2]The George Washington University, Washington, D.C., U.S.A.
(kotz@gwu.edu,hosam@gwu.edu)

Abstract. We investigate the distribution, mean value, variance and some limiting properties of an urn model of white and red balls under random multiple drawing (either with or without replacement) when the number of white and red balls added follows a schedule that depends on the number of white balls chosen in each drawing.

1 Introduction

The urn model is an abstract form of various population models, especially in biology. The basic distribution associated with the corresponding probabilistic process is the well-known Pólya distribution. A Pólya distribution can be generated by repetition of drawings from an urn. The urn starts out with τ_0 balls, of which ω_0 are white and the remaining $\tau_0 - \omega_0$ are red. A drawing is affected by selecting a ball at random from the urn, noting its color, and replacing it with c balls of the same color. For further details see, for example, Johnson, Kotz and Kemp (1992).

Received: October 2003, Revised: June 2004
Key words and phrases: Random structure, Urn model.

Table 1: The Replacement Scheme S_1.

Number of white balls chosen	0	1	2	...	j	...	k
Number of white balls added	1	0	-1	...	$-(j-1)$	...	$-(k-1)$
Number of red balls added	0	1	2	...	j	...	k

There have been many modifications of this process since the original paper of Pólya (1930), see, for example, Kotz and Balakrishnan (1997) for a recent survey. The basic models, and several modifications therein, serve as convenient tools for the probabilistic analysis of algorithms. For example, they model well random trees, which underly many sorting algorithms. Mahmoud (2003) surveys numerous applications in this area.

In the present paper we describe a modified version that has so far received attention in the literature only in the recent work of Tsukiji and Mahmoud (2001). Graph-theoretic motivation and applications in informatics for the generalization described below are lucidly discussed in Tsukiji and Mahmoud (2001), but our generalization is not restricted, as in their work, to the case of drawing only two balls in each operation. Specifically, we consider selecting k balls in each draw, observing the number W_n of white balls chosen, replacing them in the urn, and then adding or removing balls from the urn according to the scheme presented in Table 1.

This replacement scheme can be summarized as

Addition rule: *Replace the balls removed from the urn in the course of random sampling by k red balls and one white ball.*

In the scheme of Table 1 the total number of balls in the urn increases by one with each draw, hence the designation S_1. We shall briefly discuss generalization to schemes adding a total of $c \geq 1$ balls, which are to be called S_c. In the S_1 scheme, the number of balls in the urn after n draws is

$$\tau_n = \tau_0 + n. \tag{1}$$

In Section 2 we establish a formula for the expected value of the number ω_n of white balls in the urn at the conclusion of the nth draw.

In that section it is shown that the same results apply whether the sampling is with or without replacement. In later sections, on variance of ω_n and a recurrence formula for its probability distribution, we do need to make allowance for these differences in the sampling procedures. Also, it must be admitted that the without-replacement model is, in the scheme S_1, only a possibly useful approximation to the with-replacement model. This is because it is possible in sampling with replacement to observe a white ball more times than there are white balls in the urn, and the system can then require more white balls to be removed from the urn than are contained therein. The scheme would then be "untenable," as it is often called in the literature. See Balaji and Mahmoud (2003) for a characterization of tenability in the usual case of single drawing.

A scheme based on sampling with replacement can be immunized against untenability, if there are no circumstances under which the number of balls (of either color) can be reduced. Such a system could be created by increasing the number of white balls to be added in the scheme S_1 to c, where $c \geq k$.

2 Expected number of white balls

Let $\pi_n := \omega_n/\tau_n$ be the proportion of white balls in the urn after n draws. Conditioned on ω_{n-1}, the number W_n of white balls in the sample is $binomial(k, \pi_{n-1})$ in the case of sampling with replacement, and is $hypergeometric(k, \omega_{n-1}, \tau_{n-1})$ in the case of sampling without replacement. These two distributions have the same mean $k\pi_n$ (but not the same variance). The expected number of white balls chosen in the sample in the nth draw is

$$\mathbf{E}[W_n \,|\, \omega_{n-1}] = k\frac{\omega_{n-1}}{\tau_{n-1}} = k\pi_{n-1}.$$

This conditional expectation holds *whether the sampling is with or without replacement*.

Recall that ω_0, τ_0, and π_0 are deterministic parameters, and so is τ_n (in view of (1)). The number of white balls after the nth draw is equal to their number after $n-1$ draws plus the number of balls added (possibly negative) after the nth drawing. Therefore, the number of white balls in the urn follows the stochastic recurrence

$$\omega_n = \omega_{n-1} + 1 - W_n.$$

From this recurrence, it follows that the conditional expectation (given ω_{n-1}) is

$$\begin{aligned}\mathbf{E}[\omega_n \,|\, \omega_{n-1}] &= \omega_{n-1} + 1 - \mathbf{E}[W_n \,|\, \omega_{n-1}] \\ &= 1 + \Big(1 - \frac{k}{\tau_{n-1}}\Big)\omega_{n-1}, \end{aligned} \tag{2}$$

whether the sampling is with or without replacement. Taking expectations again we find

$$\mathbf{E}[\omega_n] = 1 + \Big(1 - \frac{k}{\tau_{n-1}}\Big)\mathbf{E}[\omega_{n-1}].$$

We derive next an informative rearrangement of this equation in terms of $E'_j = \mathbf{E}[\pi_j - (k+1)^{-1}]$, the expected distance between the proportion of white balls after j draws and what will turn out to be its limiting value. The last equation can be rewritten as

$$\tau_n \mathbf{E}[\pi_n] = 1 + (\tau_{n-1} - k)\mathbf{E}[\pi_{n-1}].$$

Utilizing the fact that the difference $\tau_n - \tau_{n-1}$ is 1 we write

$$\begin{aligned}\tau_n E'_n &= (\tau_{n-1} - k)E'_{n-1} + \frac{1}{k+1}(\tau_{n-1} - \tau_n + 1) \\ &= (\tau_{n-1} - k)E'_{n-1}. \end{aligned}$$

Equivalently,

$$\tau_n E'_n = \tau_{n-k-1} E'_{n-1}.$$

Unwinding this recurrence, under the initial condition $E'_0 = \pi_0 - (k+1)^{-1}$, we obtain a solution to the recurrence. This solution is most conveniently written in terms of the descending factorial

$$a^{(r)} = a(a-1)\dots(a-r+1).$$

One obtains

$$\begin{aligned}\mathbf{E}\Big[\pi_n - \frac{1}{k+1}\Big] &= \frac{(\tau_0 + n - k - 1)^{(n)}}{(\tau_0 + n)^{(n)}}\mathbf{E}\Big[\pi_0 - \frac{1}{k+1}\Big] \\ &= \frac{\tau_0^{(k+1)}}{(\tau_0 + n)^{(k+1)}}\Big(\pi_0 - \frac{1}{k+1}\Big). \end{aligned} \tag{3}$$

Clearly $\tau_0^{(k+1)}/(\tau_0+n)^{(k+1)} = O(n^{-(k+1)}) \to 0$, as $n \to \infty$. Hence,

(i)
$$\Big|\mathbf{E}[\pi_n] - \frac{1}{k+1}\Big| \le \Big|\mathbf{E}\Big[\pi_{n-1} - \frac{1}{k+1}\Big]\Big|,$$
with equality holding if $\pi_0 = \frac{1}{k+1}$;

(ii) $\lim_{n\to\infty} \mathbf{E}[\pi_n] = \frac{1}{k+1}$;

(iii) $\mathbf{E}[\pi_n] \overset{\geqq}{<} \frac{1}{k+1}$ for all n according as $\pi_0 \overset{\geqq}{<} \frac{1}{k+1}$.

The result (ii) is, of course, to be expected in view of the alternative description via the addition rule, where $(k+1)^{-1}$ is the proportion of white balls added. Since $\mathbf{E}[\pi_n] = \tau_n^{-1}\mathbf{E}[\omega_n]$, equation (3) also provides an exact formula for the average number of white balls after n draws. This takes a simple asymptotic form, namely
$$\mathbf{E}[\omega_n] = \frac{n+\tau_0}{k+1} + O\Big(\frac{1}{n^{k+1}}\Big).$$

For the immune system S_c, guaranteed to be tenable, we have $\tau_n = \tau_0 + cn$, and
$$\mathbf{E}[\omega_n] = c + \Big(1 - \frac{k}{\tau_{n-1}}\Big)\mathbf{E}[\omega_{n-1}],$$
leading, after some manipulation, to
$$\mathbf{E}\Big[\pi_n - \frac{c}{k+c}\Big] = \frac{\tau_{n-k-1}}{\tau_n}\Big(\mathbf{E}[\pi_{n-1}] - \frac{c}{k+c}\Big), \tag{4}$$
where $\tau_{n-k-1} = \tau_0 + c(n-k-1)$. Finally,

(i)
$$\Big|\mathbf{E}[\pi_n] - \frac{c}{k+c}\Big| \le \Big|\pi_0 - \frac{c}{k+c}\Big|,$$
with equality holding only when $\pi_0 = \frac{c}{k+c}$;

(ii) $\lim_{n\to\infty} \mathbf{E}[\pi_n] = \frac{c}{k+c}$;

(iii) $\mathbf{E}[\pi_n] \overset{\geqq}{<} \frac{c}{k+c}$ for all n according as $\pi_0 \overset{\geqq}{<} \frac{c}{k+c}$.

Again, the result (ii) is to be expected, since S_c is equivalent to replacing the chosen balls by c white and k red balls—namely a proportion of $c/(k+c)$ white balls. Curiously, if we take $c = k^2$, we obtain $\lim_{n\to\infty} \mathbf{E}[\pi_n] = 1 - \frac{1}{k+1}$, thus reversing the limiting proportions of white and red balls as given in the case $c = 1$.

3 Variance of the number of white balls

As was already mentioned, the results for the mean hold whether the sampling is with or without replacement. This is because even though the conditional sampling distributions are not the same, they possess the same expected value. However, to construct a formula for the variance of ω_n, we need the conditional variance of W_{n-1} (given ω_{n-1}). These conditional variances are not the same for both sampling procedures. Indeed, these conditional variances are

$$k\pi_{n-1}(1-\pi_{n-1}) \qquad \text{for sampling with replacement,} \tag{5}$$

and

$$k\pi_{n-1}(1-\pi_{n-1})\frac{\tau_{n-1}-k}{\tau_{n-1}-1} \qquad \text{for sampling without replacement.} \tag{6}$$

We employ the standard conditional variance formula

$$\mathbf{Var}[\omega_n] = \mathbf{E}\Big[\mathbf{Var}[\omega_n \,|\, \omega_{n-1}]\Big] + \mathbf{Var}\Big[\mathbf{E}[\omega_n \,|\, \omega_{n-1}]\Big].$$

Now

$$\mathbf{Var}[\omega_n \,|\, \omega_{n-1}] = k\theta_n\pi_{n-1}(1-\pi_{n-1)},$$

where, by the formulas (5) and (6), we have

$$\theta_n = \begin{cases} 1, & \text{for sampling with replacement;} \\ \frac{\tau_{n-1}-k}{\tau_{n-1}-1}, & \text{for sampling without replacement,} \end{cases}$$

and by (2)

$$\begin{aligned} \mathbf{Var}\Big[\mathbf{E}[\omega_n \,|\, \omega_{n-1}]\Big] &= \mathbf{Var}\Big[1 + \Big(1 - \frac{k}{\tau_{n-1}}\Big)\omega_{n-1}\Big] \\ &= \Big(1 - \frac{k}{\tau_{n-1}}\Big)^2 \mathbf{Var}[\omega_{n-1}]. \end{aligned}$$

So,

$$\mathbf{Var}[\omega_n] = k\theta_n\mathbf{E}[\pi_{n-1}(1-\pi_{n-1})] + \Big(1 - \frac{k}{\tau_{n-1}}\Big)^2 \mathbf{Var}[\omega_{n-1}],$$

or equivalently,

$$\mathbf{Var}[\pi_n] = \frac{k\theta_n}{\tau_n^2}\mathbf{E}[\pi_{n-1}(1-\pi_{n-1})] + \frac{(\tau_{n-1}-k)^2}{\tau_n^2}\mathbf{Var}[\pi_{n-1}]$$

$$= \frac{k\theta_n}{\tau_n^2}\mathbf{E}[\pi_{n-1}](1-\mathbf{E}[\pi_{n-1}]) + \frac{(\tau_{n-1}-k)^2 - k\theta_n}{\tau_n^2}\mathbf{Var}[\pi_{n-1}].$$

The first term in the last equation is $O(n^{-2})$. Consequently,

$$\mathbf{Var}[\pi_n] = O\Big(\frac{1}{n}\Big).$$

The orders of magnitude of the mean and variance give us a concentration law:

$$P\Big(|\pi_n - \mathbf{E}[\pi_n]| \geq \varepsilon\Big) \leq \frac{\mathbf{Var}[\pi_n]}{\varepsilon^2} = O\Big(\frac{1}{n}\Big),$$

for any fixed $\varepsilon > 0$. In other words,

$$\pi_n \to \frac{1}{k+1}, \qquad \text{in probability.}$$

4 A recursion formula for the distribution

In this section we assume tenability. For convenience we use the abbreviation

$$P_n(\omega) := P(\omega_n = \omega).$$

The event $\{\omega_n = \omega\}$ can arise in the following mutually exclusive ways:

$$\{\omega_{n-1} = \omega + j - 1\} \cap \{W_n = j\}, \qquad \text{for } \; j = 0, 1, \ldots, k.$$

Hence,

$$\begin{aligned} P_n(\omega) &= \sum_{j=0}^{k} P_{n-1}(\omega + j - 1)\, P(W_n = j \,|\, \omega_{n-1} = \omega + j - 1) \\ &= \sum_{j=0}^{k} a_j(\omega) P_{n-1}(\omega + j - 1), \end{aligned} \tag{7}$$

where in the case of sampling with replacement

$$a_j(\omega) = \binom{k}{j}\Big(\frac{\omega + j - 1}{\tau_0 + n - 1}\Big)\Big(1 - \frac{\omega + j - 1}{\tau_0 + n - 1}\Big),$$

while in the case of sampling without replacement

$$a_j(\omega) = \binom{k}{j} \frac{(\omega + j - 1)^{(j)} (\tau_0 + n - \omega - j)^{(k-j)}}{(\tau_0 + n - 1)^{(k)}}.$$

The recurrence is to be solved under the boundary condition

$$P_0(\omega) = \begin{cases} 1, & \text{for } \omega = \omega_0; \\ 0, & \text{oherwise.} \end{cases}$$

Note that for the scheme S_1, if $\omega_{n-1} = 0$, then necessarily $\omega_n = 1$ (for the system S_c, $\omega_n = c$). Explicit values of $P_n(\omega)$ can be obtained from (7), in specific cases, by straightforward numerical calculation.

5 Multivariate extensions

A direct multivariate extension of the urn model described in this paper arises when there are balls of several (say p) different colors originally in the urn and the replacement set also contains balls of these (and possibly other) colors. Each drawing is affected by taking a random sample of size k from the urn, and adding a replacement set containing γ_i balls of color i, for $i = 1, \ldots, p$. The replacement set is of course of total size $\gamma := \sum_{i=1}^{p} \gamma_i$ that is no less than k. A modified form of equation (4) and the three consequent properties (i)–(iii) will apply to any particular color (though now $\tau_n = \tau_0 + (\gamma - k)n$). An interesting track to pursue is determining the joint distribution of the number of balls of different colors.

Another possibility envisions a series (cascade) of urns in which the selected random sample from an urn is discarded in the next urn in the series, with each urn having its own replacement set. (A "circular" cascade can be constructed by having the random sample from the last urn in the series be deposited in the first.) So, other variants of the model presented are possible, but this is for another time!

References

[1] Balaji, S. and Mahmoud, H. (2003), Problem 762. The College Mathematics Journal, **34**, 405–406.

[2] Johnson, N., Kotz, S., and Kemp, A. (1992), Univariate Discrete Distributions. Second Edition, New York: Wiley.

[3] Kotz, S. and Balakrishnan, N. (1997), Advances in urn models during the past two decades. In Advances in Combinatorial Methods and Applications to Probability and Statistics, **49**, 203–257. Birkhäuser, Boston.

[4] Mahmoud, H. (2003), Urn models and connections to random trees: A review. Journal of the Iranian Statistical Society, **2**, 53–114.

[5] Pólya, G. (1930), Sur quelques points de la théorie des probabilités. Annals of the Institute of Henri Poincaré, **1**, 117–161.

[6] Tsukiji, T. and Mahmoud, H. (2001), A limit law for outputs in random circuits. Algorithmica, **31**, 403–412.