Ann Inst Stat Math (2017) 69:1129–1139
DOI 10.1007/s10463-016-0578-5

On coupon collector's and Dixie cup problems under fixed and random sample size sampling schemes

James C. Fu[1] · **Wan-Chen Lee**[2]

Received: 21 April 2016 / Revised: 13 June 2016 / Published online: 9 August 2016

Abstract Suppose an urn contains m distinct coupons, labeled from 1 to m. A random sample of k coupons is drawn without replacement from the urn, numbers are recorded and the coupons are then returned to the urn. This procedure is done repeatedly and the sample sizes are independently identically distributed. Let W be the total number of random samples needed to see all coupons at least l times ($l \geq 1$). Recently, for $l = 1$, the approximation for the first moment of the random variable W has been studied under random sample size sampling scheme by Sellke (Ann Appl Probab, 5:294–309, 1995). In this manuscript, we focus on studying the exact distributions of waiting times W for both fixed and random sample size sampling schemes given $l \geq 1$. The results are further extended to a combination of fixed and random sample size sampling procedures.

Keywords Coupon collector's problems · Dixie cup problems · Finite Markov chain imbedding

1 Introduction

Pólya (1930) showed that the average waiting time to see all the distinct coupons at least once ($l = 1$) is equivalent to the average waiting time of the followingurn

This work was partially support by Grant A-9216 of the Natural Science and Engineering Research Council of Canada.

✉ James C. Fu
james.fu@umanitoba.ca

Wan-Chen Lee
wanchen.lee@canada.ca

[1] Department of Statistics, University of Manitoba, Winnipeg R3T 2N2, Canada

[2] Healthy Environments and Consumer Safety Branch, Health Canada, Ottawa K1A 0K9, Canada

problem. It starts with m white balls in an urn. A ball is drawn from the urn with replacement. If the observed ball is white, it will be colored in red and returned to the urn. Otherwise, the selected ball will simply be returned to the urn without changing its color. Then the average number of drawings until all m balls are in red equals the average number of trials to see all the distinct coupons at least once. This problem is referred as a waiting time problem by Feller (1957). It was pointed out by Ivchenko (1998) that the coupon collector's problem is closely related to "number random allocations" (see Markoff 1912). Further the coupon collector's problem can also be viewed as a special case of Dixie cup problem or urn problem, see Newman and Shepp (1960) and Erdös and Rényi (1961). Since then it has attracted many mathematicians, probabilitists, computer scientists, and statisticians for its theoretical challenge and various applications.

For $k = 1$, let $W_{m,l}$ be the number of random samples needed to see all m coupons at least l times. For $l \geq 2$, it is often referred as Dixie cup problem (see Newman and Shepp 1960). They have shown, as $m \to \infty$,

$$\mathbf{E}W_{m,l} = m \log m + (l-1)m \log\log m + mC_l + o(m), \tag{1}$$

where C_l is a constant, depending on l. For $l = 1$, Eq. (1) yields the classical result $\mathbf{E}W_{m,1} \sim m \log m + mC_1$. For given l, Erdös and Rényi (1961) obtained the limiting distribution for the random variable $W_{m,l}$

$$\lim_{m\to\infty} P\left(\frac{W_{m,l}}{m} < \log m + (l-1)\log\log m + x\right) = \exp\left(-\frac{\mathrm{e}^{-x}}{(l-1)!}\right), \tag{2}$$

for every real x. The problem has been extended to various directions, for example to name a few, Adler and Ross (2001), Erdös and Rényi (1961), Dawkins (1991) and Klaassen (1994).

Let Z_i be a sequence of *i.i.d.* random sample size variables defined on the support $\{1, 2, \cdots, K : K \leq m\}$, having common distribution $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$. For $\mu_1 = 1$ and $\mu_i = 0$ for $i = 2, \ldots, K$, this setting is the classical coupon collector's problem and $\mathbf{E}W_{m,1} = m\sum_{i=1}^{m} i^{-1}$. For $\mu_k = 1$, $1 \leq k \leq K$, Pólya (1930) made use of inclusion–exclusion argument to obtain the exact formula for $\mathbf{E}W_{m,1}$ for a fixed sample size sampling scheme S_k. The formula proposed is somewhat difficult to evaluate, even for moderate m. The good approximation for $\mathbf{E}W_{m,1}$ is needed for moderate or large m. For $l = 1$ and under random sample size sampling scheme $S_{\boldsymbol{\mu}}$, Sellke (1995) provided the following formula for approximating the expected waiting time $\mathbf{E}W_{m,1}(S_{\boldsymbol{\mu}})$ by using Wald's identity and Markov chain coupling:

$$\mathbf{E}W_{m,1}(S_{\boldsymbol{\mu}}) \approx \frac{\sum_{i=1}^{m-1} \dfrac{1}{(m-i)}}{\sum_{i=1}^{m-1} \dfrac{p_i}{(m-i)}} + \frac{\sum_{i=1}^{m-1} \left(\dfrac{p_i}{(m-i)} \sum_{j=1}^{i} \dfrac{1}{(m-j+1)}\right)}{\left(\sum_{i=1}^{m-1} \dfrac{p_i}{(m-i)}\right)^2}, \tag{3}$$

where $p_i = P(Z > i)$. Sellke (1995) also pointed out that the above approximation performs very well when the support $\{1, \ldots, K\}$ has a high probability in the

range $\{1, \ldots, m/2\}$. Ivchenko (1998) gave the exact formulae for $\mathbf{E}W_{m,1}(S_{\boldsymbol{\mu}})$ and $\mathrm{Var}(W_{m,1}(S_{\boldsymbol{\mu}}))$ by using Polya's results for fixed sample size cases. Recently, Johnson and Sellke (2010) generalized the partial fraction technique developed by Pólya (1930) for approximating $\mathbf{E}W_{m,1}(S_{\boldsymbol{\mu}})$. They have shown their approximation agrees with Eq. (3).

In this manuscript, we use Polya's coloring technique together with finite Markov chain imbedding (FMCI) method (see Fu and Koutras 1994 and Fu and Lou 2003) to study the following events:

(i) the exact distributions of waiting times for coupon collector's and Dixie cup problems ($l \geq 1$) under fixed sample size sampling schemes S_k and random sample size sampling schemes $S_{\boldsymbol{\mu}}$,
(ii) the waiting time distributions to see $m'(m' < m)$ distinct coupons or to see $\tilde{m}(\tilde{m} < m)$ specified distinct coupons at least once (or l times), and
(iii) the relationship among fixed sample size, random sample size, and mixed sample size sampling schemes.

As a direct consequence from the method of FMCI, we have also obtained the mean, variance and moment generating function for the waiting time $W_{m,l}$. Examples and numerical results are also provided to illustrate the theoretical results.

2 Waiting time distributions

2.1 For fixed and random sample size sampling schemes with $l = 1$ (coupon collector's problem)

Let S_k be the fixed sample size $k(k = 1, 2, \ldots, m)$ sampling scheme. The waiting time random variable $W_{m,1}(S_k)$, the number of random samples required to observe all the m distinct coupons at least once ($l = 1$), is finite Markov chain imbeddable in the following sense.

Theorem 1 *Given m, $l = 1$, and sampling scheme S_k, there exists a homogeneous Markov chain $\{Y_t\}$ defined on the state space $\Omega = \{0, 1, \ldots, m-1, m = \alpha\}$ with α as an absorbing state and the transition probability matrix having the form,*

$$\boldsymbol{M}_{m,1}(S_k) = [Pij(S_k)] = \begin{array}{c} \\ \Omega\backslash\alpha \\ \alpha \end{array} \begin{array}{c} \begin{array}{cc} \Omega\backslash\alpha & \alpha \end{array} \\ \left[\begin{array}{c|c} \boldsymbol{N}_{m,1}(S_k) & \boldsymbol{C}_{m,1}(S_k) \\ \hline \mathbf{0} & 1 \end{array}\right] \end{array} \tag{4}$$

where

$$p_{ij}(S_k) = \begin{cases} \dfrac{\binom{i}{k-j+i}\binom{m-i}{j-i}}{\binom{m}{k}} & \text{for } 0 \leq i \leq j \leq m \text{ and } j - 1 \leq \min(k, m-i), \\ 0 & \text{otherwise}, \end{cases} \tag{5}$$

then, given initial distribution $\boldsymbol{\xi}_0 = (1, 0, \ldots, 0)$, we have

$$P(W_{m,1}(S_k) > n | \boldsymbol{\xi}_0) = \boldsymbol{\xi}_0 \boldsymbol{N}_{m,1}^n(S_k)\mathbf{1}'. \tag{6}$$

Proof It is easy to see that, for given $Y_t = i$ (the number of red balls in the urn), the random variable X, the number of white balls in a sample of size k sampling without replacement from an urn containing m balls with $m - i$ white and i red balls, has a hypergeometric distribution $HG(m, i, k)$. Hence, the transition probabilities

$$p_{ij}(S_k) = P(Y_{t+1} = j | Y_t = i) = P(X = j - i)$$

are defined by the Eq. (5) and also independent of t. The imbedded chain $\{Y_t\}$ is a homogeneous Markov chain with the state space $\Omega = \{0, 1, \ldots, m-1, m = \alpha\}$ and state m as an absorbing state α. Throughout the remainder of the manuscript, α is used as an absorbing state. Note also that the two events are equivalent in the following way:

$$\{W_{m,1}(S_k) > n\} \quad \text{if and only if} \quad \{Y_1 \neq \alpha, \ldots, Y_n \neq \alpha\} \tag{7}$$

for $n = 1, \ldots, \infty$. Given an initial distribution $\boldsymbol{\xi}_0$ and the above statement (7), we have

$$P(W_{m,1}(S_k) > n | \boldsymbol{\xi}_0) = P(Y_1 \neq \alpha, \ldots, Y_n \neq \alpha | \boldsymbol{\xi}_0) = \boldsymbol{\xi}_0 \boldsymbol{N}^n_{m,1}(S_k) \mathbf{1}'.$$

This complete the proof of Eq. (6). □

By the same token, the following theorem shows that the waiting time random variable $W_{m,1}(S_{\boldsymbol{\mu}})$ based on a random sample size sampling scheme is also finite Markov chain imbeddable. Further, the relationship between transition probability matrix $\boldsymbol{M}_{m,1}(S_{\boldsymbol{\mu}})$ and the transition probability matrices $\boldsymbol{M}_{m,1}(S_k)$, $k = 1, \ldots, K$ for fixed sample size sampling schemes are developed.

Theorem 2 *Given m, $l = 1$, and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$, the waiting time random variable $W_{m,1}(S_{\boldsymbol{\mu}})$ is finite Markov chain imbeddable in the sense that there exists a homogeneous Markov chain $\{Y_t\}$ defined on the state space $\Omega = \{0, 1, \ldots, m = \alpha\}$ with transition probability matrix*

$$\boldsymbol{M}_{m,1}(S_{\boldsymbol{\mu}}) = [\ p_{ij}(S_{\boldsymbol{\mu}})\ ] = \left[\begin{array}{c|c} \boldsymbol{N}_{m,1}(S_{\boldsymbol{\mu}}) & \boldsymbol{C}_{m,1}(S_{\boldsymbol{\mu}}) \\ \hline \mathbf{0} & 1 \end{array}\right],$$

where

$$p_{ij}(S_{\boldsymbol{\mu}}) = \sum_{k=1}^{K} \mu_k \frac{\binom{i}{k-j+i}\binom{m-i}{j-i}}{\binom{m}{k}} I(m, i, k) \tag{8}$$

and

$$I(m, i, k) = \begin{cases} 1 & \text{for } 0 \leq i \leq j \leq m \text{ and } j - 1 \leq \min(k, m - i), \\ 0 & \text{otherwise.} \end{cases}$$

Then, given initial distribution $\boldsymbol{\xi}_0 = (1, 0, \ldots, 0)$,

$$P(W_{m,1}(S_{\boldsymbol{\mu}}) > n|\boldsymbol{\xi}_0) = \boldsymbol{\xi}_0 \boldsymbol{N}^n_{m,1}(S_{\boldsymbol{\mu}})\mathbf{1}'. \tag{9}$$

Proof Note that the Eq. (8) is a direct consequence of the following equation and equation (5)

$$p_{ij}(S_{\boldsymbol{\mu}}) = \sum_{k=1}^{K} \mu_k P(Y_{t+1} = j|Y_t = i, Z_t = k).$$

This result (9) follows same arguments as the proof of Theorem 1. This completes the proof. □

In view of above formulae (5) and (8), it can be clearly seen that the essential transition probability matrix associated with the random sample size sampling scheme is a weighted sum of essential transition probability matrices of fixed sample size sampling schemes; i.e.,

$$\boldsymbol{N}_{m,1}(S_{\boldsymbol{\mu}}) = \sum_{k=1}^{K} \mu_k \boldsymbol{N}_{m,1}(S_k).$$

Furthermore, let $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{iK}), i = 1, \ldots, d$ be d random sample size sampling schemes and $S_{\boldsymbol{\beta}}, \boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)$ be a linear combination of d random sample size sampling schemes, then we have

$$\boldsymbol{N}_{m,1}(S_{\boldsymbol{\beta}}) = \sum_{i=1}^{d} \beta_i \boldsymbol{N}_{m,1}(S_{\boldsymbol{\mu}_i}) = \sum_{i=1}^{d}\sum_{k=1}^{K} \beta_i \mu_{ik} \boldsymbol{N}_{m,1}(S_k).$$

Let S represent a generic sampling scheme with either fixed sample size S_k or random sample size $S_{\boldsymbol{\mu}}$, or mixed random sample sizes $S_{\boldsymbol{\beta}}$. As we have proved above that the waiting time random variable $W_{m,1}(S)$ is finite Markov chain imbeddable. With some simple algebra, it follows directly from the definitions and equations (5) and (8) that the random variable $W_{m,1}(S)$ has (see Fu and Lou 2003)

(i) the moment generating function

$$\varphi_{W_{m,1}(u)} = 1 + (\mathrm{e}^u - 1)\boldsymbol{\xi}_0(\boldsymbol{I} - \mathrm{e}^u \boldsymbol{N}_{m,1}(S))^{-1}\mathbf{1}' \quad \text{for } u \in R, \tag{10}$$

(ii) the mean

$$\mathbf{E}W_{m,1}(S) = \boldsymbol{\xi}_0(\boldsymbol{I} - \boldsymbol{N}_{m,1}(S))^{-1}\mathbf{1}', \tag{11}$$

and

(iii) the second moment

$$\mathbf{E}[W_{m,1}(S)]^2 = \boldsymbol{\xi}_0(\boldsymbol{I} + \boldsymbol{N}_{m,1}(S))(\boldsymbol{I} - \boldsymbol{N}_{m,1}(S))^{-2}\mathbf{1}'. \tag{12}$$

Note that these characters of the distribution of $W_{m,1}(S)$ are all in terms of essential transition probability matrix $\mathbf{N}_{m,1}(S)$ in a very simple form. This is due to the fact that the imbedded Markov chain $\{Y_t\}$ is completely characterized by its transition probability matrix.

For large n, the tail probability $P(W_{m,1}(S) > n|\boldsymbol{\xi}_0)$ can be approximated by the largest eigenvalue of essential transition probability matrix in the following way. Let $1 > \gamma_1 \geq |\gamma_2| \geq \cdots \geq |\gamma_q|$ be the q eigenvalues of the essential transition probability matrix $\boldsymbol{N}_{m,l}(S)$ and $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \ldots, \boldsymbol{\eta}_q$ be the eigenvectors corresponding to the eigenvalues $\gamma_1, \gamma_2, \ldots, \gamma_q$, respectively. It follows that

$$P(W_{m,1}(S) > n|\boldsymbol{\xi}_0) = ab\exp\{n[\log\gamma_1 + o(1)]\}, \tag{13}$$

where $a = \boldsymbol{\eta}_1\mathbf{1}'$ and $b = \boldsymbol{\xi}_0\boldsymbol{\eta}_1'$. Based on our experience, the numerical approximations based on Eq. (13) performs very well. For details and proofs, see Fu and Johnson (2009).

For a special case of $k = 1$, it follows from equation (4) the transition probability matrix $\boldsymbol{M}_{m,1}(S_1)$ of the imbedded Markov chain $\{Y_t\}$ defined on the state space $\Omega = \{0, 1, \ldots, m\}$ corresponding to the waiting time random variable $W_{m,1}(S_1)$ has the form

$$\boldsymbol{M}_{m,1}(S_1) = [\ p_{ij}(S_1)\ ] = \left[\begin{array}{c:c} \boldsymbol{N}_{m,1}(S_1) & \boldsymbol{C} \\ \hdashline \mathbf{0} & 1 \end{array}\right], \tag{14}$$

where for $0 \leq i \leq j \leq m$,

$$p_{ij}(S_1) = \begin{cases} i/m & \text{if } \ j = i, \\ (m-i)/m & \text{if } \ j = i+1, \\ 0 & \text{otherwise.} \end{cases} \tag{15}$$

It follows from Eqs. (11), (14) and (15), we have the classic result

$$\mathbf{E}W_{m,1}(S_1) = \boldsymbol{\xi}_0(\boldsymbol{I} - \boldsymbol{N}_{m,1}(S_1))^{-1}\mathbf{1}' = m\left(1 + \frac{1}{2} + \cdots \frac{1}{m}\right).$$

For $m \to \infty$, we have

$$\mathbf{E}W_{m,1}(S_1) = m\log m + mC_1 + o(m),$$

where C_1 is so-called Euler's constant.

In the following, we provide two technique remarks for the direct extensions of above results.

Remark 1 With simple modification on the state space, the above result can be extended to the waiting time random variable $W_{m,1}(S_1 : m')$ to see $m'(1 \le m' \le m)$ distinct coupons at least once. In this case, the corresponding imbedded Markov chain $\{Y_t\}$ is defined on the state space $\Omega = \{0, 1, \ldots, m' - 1, m' = \alpha\}$ and having the transition probabilities, for $0 \le i \le j \le m' \le m$,

$$p_{ij}(S_1 : m') = \begin{cases} i/m & \text{if } j = i, i \ne m', \\ (m-i)/m & \text{if } j = i+1, i \ne m', \\ 1 & \text{if } i = j = m', \\ 0 & \text{otherwise.} \end{cases}$$

Remark 2 Let $W_{m,1}(S_1 : \tilde{m})$ be the waiting time to see specifically specified $\tilde{m}(1 \le \tilde{m} \le m)$ distinct coupons at least once. With simple modification of Polya's arguments, it is easy to see the random variable $W_{m,1}(S_1 : \tilde{m})$ is also finite Markov chain imbeddable and the imbedded Markov chain $\{Y_t\}$ is defined on the state space $\Omega = \{0, 1, \ldots, \tilde{m} - 1, \tilde{m} = \alpha\}$ and having the transition probabilities, for $0 \le i \le j \le \tilde{m} \le m$,

$$p_{ij}(S_1 : m') = \begin{cases} (m - \tilde{m} + i)/m & \text{if } j = i, i \ne \tilde{m}, \\ (\tilde{m} - i)/m & \text{if } j = i+1, i \ne \tilde{m}, \\ 1 & \text{if } i = j = \tilde{m}, \\ 0 & \text{otherwise.} \end{cases}$$

2.2 Fixed and random sample size sampling schemes with $l \ge 2$ (Dixie cup problem)

For $l = 1$, Pólya (1930) used two colored balls white and red technique to solve the coupon collector's problem. For $l \ge 2$, the technique can be extended to involve $l + 1$ colors, say colors $c_0, c_1, \ldots, c_l$. Let us consider an urn containing m balls with m_i balls of color c_i for $i = 0, 1, \ldots, l$, $0 \le m_i \le m$ and $\sum_{i=0}^{l} m_i = m$. For $i = 0, 1, \ldots, l-1$, m_i means the number of the balls that have been observed exactly i times and m_l observed at least l times. A random sample of $k(1 \le k \le m)$ balls is drawn without replacement from the urn which contains $k_0, k_1, \ldots, k_l$ balls of colors $c_0, c_1, \ldots, c_l$, respectively, where $0 \le k_i \le m_i$ and $k_0 + k_1 + \cdots + k_l = k$. Then k_i balls of color c_i are colored in color c_{i+1} for $i = 0, 1, \ldots, l-1$ and k_l balls of color c_l are kept their color. It follows after the sampling that the urn contains $(m_0 - k_0)$ balls of color c_0, $(m_1 - k_1 + k_0)$ balls of color c_1, …, $(m_{l-1} - k_{l-1} + k_{l-2})$ balls of color c_{l-1} and $(m_l + k_{l-1})$ balls of color c_l. After that, all k balls are then returned to the urn. The procedure is repeated until all balls in the urn having the same color c_l. The waiting time random variable $W_{m,l}(S_k)$ associated with above sampling scheme is homogeneous finite Markov chain imbeddable. The homogeneous imbedded Markov chain $\{Y_t\}$, induced by the S_k sampling scheme, has the following structure.

For fixed m and l, we define the state space Ω of imbedded Markov chain $\{Y_t\}$:

$$\Omega = \left\{ \boldsymbol{\omega} = (m_1, m_2, \ldots, m_l) : 0 \le m_i \le m \ \text{ for } \ i = 1, \ldots, l \ \text{ and } \ \sum_{i=0}^{l} m_i = m \right\}, \tag{16}$$

where $m_0 = m - \sum_{i=1}^{l} m_i$, $(0, 0, \cdots, 0)$ stands as an initial state and $(0, \ldots, 0, m)$ an absorbing state α. For $\boldsymbol{\omega} = (m_1, m_2, \ldots, m_l) \in \Omega$, we define

$$\mathcal{A}_{\boldsymbol{\omega},k} = \left\{ \boldsymbol{k} = (k_1, k_2, \ldots, k_l) : 0 \le k_i \le m_i \ \text{ for } \ i = 0, 1, \ldots, l \ \text{ and } \ \sum_{i=0}^{l} k_i = k \right\}. \tag{17}$$

Given $Y_t = \boldsymbol{\omega}$, Y_{t+1} is defined the following way: for every $\boldsymbol{k} = (k_1, k_2, \ldots, k_l) \in \mathcal{A}_{\boldsymbol{\omega},k}$

$$\begin{aligned} Y_{t+1} = \langle \boldsymbol{\omega}, \boldsymbol{k} \rangle = (&m_1 - k_1 + k_0, m_2 - k_2 + k_1, \\ &\ldots, m_{l-1} - k_{l-1} + k_{l-2}, m_l + k_{l-1}). \end{aligned} \tag{18}$$

Note that, given $\boldsymbol{\omega} \in \Omega$, the random vector $\boldsymbol{k}$ follows a Multi-Hypergeometric distribution (MHG$(m, \boldsymbol{\omega}, \boldsymbol{k})$) hence the transition probabilities are:

$$p_{\boldsymbol{\omega},\boldsymbol{\omega}'}(S_k) = \begin{cases} \dfrac{\prod_{i=0}^{l} \binom{m_i}{k_i}}{\binom{m}{k}} & \text{if } \ \boldsymbol{\omega}' = \langle \boldsymbol{\omega}, \boldsymbol{k} \rangle \text{ and } \boldsymbol{k} \in \mathcal{A}_{\boldsymbol{\omega},k}, \\ 0 & \text{otherwise.} \end{cases} \tag{19}$$

Denote $\boldsymbol{M}_{m,l}(S_k)$ and $\boldsymbol{N}_{m,l}(S_k)$ be the transition probability matrix and essential transition probability matrix induced by the Eq. (19), respectively. Then the following theorem holds.

Theorem 3 *The random variable $W_{m,l}(S_k)$ is finite Markov chain imbeddable and*

$$P(W_{m,l}(S_k) > n | \boldsymbol{\xi}_0) = \boldsymbol{\xi}_0 \boldsymbol{N}_{m,l}^n(S_k) \mathbf{1}'. \tag{20}$$

For a random sample size sampling scheme $S_{\boldsymbol{\mu}}$, the above Eq. (20) holds with

$$\boldsymbol{N}_{m,l}(S_{\boldsymbol{\mu}}) = \sum_{k=1}^{K} \boldsymbol{\mu}_k \boldsymbol{N}_{m,l}(S_k)$$

or

$$p_{\boldsymbol{\omega},\boldsymbol{\omega}'}(S_{\boldsymbol{\mu}}) = \sum_{k=1}^{K} \boldsymbol{\mu}_k \frac{\prod_{i=0}^{l} \binom{m_i}{k_i}}{\binom{m}{k}}.$$

To illustrate above theoretical results, we provide the following detailed example of $l = 2$ with fixed sample size sampling scheme S_1.

Example 1 Suppose three distinct balls are available and select one ball at a time, in average, how many random samples are needed to take in order to observe all three distinct balls at least twice?

The waiting time $W_{3,2}(S_1)$ is Markov chain imbeddable. The imbedded Markov chain $\{Y_t\}$ has the state space

$$\Omega = \{(0,0), (1,0), (0,1), (2,0), (1,1), (0,2), (3,0), (2,1), (1,2), (0,3) = \alpha\}.$$

Given $\boldsymbol{\omega} \in \Omega$, the 10 sets $\mathcal{A}_{\boldsymbol{\omega},k}$ are $\mathcal{A}_{(0,0),1} = \{(0,0)\}$, $\mathcal{A}_{(1,0),1} = \{(0,0), (1,0)\}$, $\mathcal{A}_{(0,1),1} = \{(0,0), (0,1)\}$, $\cdots$, $\mathcal{A}_{(1,1),1} = \{(0,0), (1,0), (0,1)\}$, $\cdots$, $\mathcal{A}_{(1,2),1} = \{(1,0), (0,1)\}$ and $\mathcal{A}_{(0,3),1} = \{(0,1)\}$. For every $\boldsymbol{\omega} \in \Omega$, the $\boldsymbol{k} \in \mathcal{A}_{\boldsymbol{\omega},1}$ yields the transition probabilities

$$p_{\boldsymbol{\omega},\boldsymbol{\omega}'}(S_1) = \begin{cases} \dfrac{\prod_{i=0}^{2} \binom{m_i}{k_i}}{3} & \text{if } \boldsymbol{\omega}' = \langle \boldsymbol{\omega}, \boldsymbol{k} \rangle \text{ and } \boldsymbol{k} \in \mathcal{A}_{\boldsymbol{\omega},1}, \\ 0 & \text{otherwise.} \end{cases}$$

The transition probability matrix $\boldsymbol{M}_{3,2}(S_1)$ has the form

$$\boldsymbol{M}_{3,2}(S_1) = \begin{array}{c} \\ (0,0) \\ (1,0) \\ (0,1) \\ (2,0) \\ (1,1) \\ (0,2) \\ (3,0) \\ (2,1) \\ (1,2) \\ (0,3) \end{array} \left[\begin{array}{cccccccccc} (0,0) & (1,0) & (0,1) & (2,0) & (1,1) & (0,2) & (3,0) & (2,1) & (1,2) & (0,3) \\ \hline 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 2/3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 0 & 2/3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2/3 & 0 & 1/3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/3 & 1/3 & 0 & 1/3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/3 & 2/3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2/3 & 1/3 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right]$$

$$= \left[\begin{array}{c|c} \boldsymbol{N}_{3,2}(S_1) & \boldsymbol{C}_{3,2}(S_1) \\ \hline \boldsymbol{0} & 1 \end{array}\right].$$

Then the expected number of random samples required to observe all three balls at least twice is

$$\mathbf{E}W_{3,2}(S_1) = \boldsymbol{\xi}_0(\boldsymbol{I} - \boldsymbol{N}_{3,2}(S_1))^{-1}\mathbf{1}' = 9.6398.$$

3 Numerical results and discussion

First, we would like to present some numerical results on distributions and expectations of both fixed sample size and random sample size sampling schemes in the following Tables 1 and 2. The computations were done on a DELL desktop using the software MATLAB®. The CPU time to complete the calculations is from few seconds for $l = 1$ and $m = 10$ to less than a minute for $l = 3$ and $m = 20$. In the following we would like to provide several technique comments on our method and numerical approximations.

Table 1 Cumulative distributions of $W_{m,l}(S_1)$ for $l = 1, 2, 3$ and their first moments and standard deviations

n	$P(W_{m,1}(S_1) \le n)$		$P(W_{m,2}(S_1) \le n)$		$P(W_{m,3}(S_1) \le n)$	
	$m = 10$	$m = 20$	$m = 10$	$m = 20$	$m = 10$	$m = 20$
20	0.2147	0.0000	0.0000	0.0000	0.0000	0.0000
25	0.4366	0.0000	0.0081	0.0000	0.0000	0.0000
35	0.7675	0.0098	0.2079	0.0000	0.0024	0.0000
40	0.8581	0.0359	0.3858	0.0000	0.0284	0.0000
45	0.9147	0.0875	0.5561	0.0000	0.1093	0.0000
50	0.9491	0.1642	0.6943	0.0000	0.2433	0.0000
75	0.9963	0.6379	0.9657	0.0706	0.8453	0.0001
100	0.9997	0.8865	0.9968	0.4444	0.9806	0.0465
250	1.0000	0.9999	1.0000	0.9992	1.0000	0.9946
300	1.0000	1.0000	1.0000	0.9999	1.0000	0.9994
500	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$\mathbb{E}(W_{m,l}(S_1))$	29.2897	71.9548	46.2296	108.6974	61.3661	141.1043
$\sigma(W_{m,l}(S_1))$	11.2110	23.8015	13.3007	27.6896	14.8737	30.6503

Table 2 Expectations of $W_{m,1}(\mu)$

m	$Z \sim \mathrm{U}(1,5)$[a]	$Z - 1 \sim \mathrm{B}(4, \frac{1}{2})$[b]
10	8.7424	8.8933
15	15.4465	15.6048
20	22.7402	22.9075
25	30.4728	30.6484
30	38.5537	38.7365
40	55.5384	55.7336

[a] Stands as a discrete uniform distribution on $\{1, 2, 3, 4, 5\}$
[b] Stands as a binomial distribution with parameters $n = 4$ and $p = 1/2$

For large l, the sizes of the state space Ω and essential transition probability matrix of imbedded Markov chain increase very fast as $m \to \infty$. In this case, the computation for distribution and expectation of the waiting time W can be problematic.

For a fixed, random, or mixed sample size sampling scheme, the random variable $W_{m,l}(S)$ is finite Markov chain imbeddable. It follows from same arguments of Eq. (13) the tail probability $P(W_{m,l}(S) > n|\boldsymbol{\xi}_0)$ can be approximated in terms of the largest eigenvalue of essential transition probability matrix of imbedded Markov chain.

Based on our experiences of computations, given l, the approximation in Eq. (1) for the expectation $\mathbf{E}W_{m,l}(S_1)$ proposed by Newman and Shepp (1960) does not perform well in the sense that

$$\frac{1}{m}\left[\mathbf{E}W_{m,l}(S_1) - m\log m - (l-1)\log\log m\right] \to C_l, \quad \text{as} \quad m \to \infty,$$

converges to C_l very slowly. For a random sample size sampling scheme $S_{\boldsymbol{\mu}}$ when the support of $\boldsymbol{\mu}$ has a high probability in the range of $[m/2, K]$, our numerical results

confirm Sellke's statement that Eq. (3) for approximating expectation $\mathbf{E}W_{m,1}(S_{\boldsymbol{\mu}})$ does not perform well.

With some modifications, the method proposed in this manuscript could be extended to handle many urn and birthday related problems. The method could also be extended to unequal probabilities model.

References

Adler, I., Ross, S. M. (2001). The coupon subset collection problem. *Journal of Applied Probability*, *38*, 737–746.

Dawkins, B. (1991). Siobhan's problem: the coupon collector revisited. *The American Statistian*, *45*, 76–82.

Erdös, P., Rényi, A. (1961). On a classical problem of probability theory. *Magyar Tudományos Akadémia Matematikai Kutató Intézetének Közleményei*, *6*, 215–220.

Feller, W. (1957). *An Introduction to Probability Theory and its Applications* (Vol. I). New York, Chichester, Brisbane: Wiley.

Fu, J. C., Johnson, B. C. (2009). Approximate Probability for Runs and Patterns in I.I.D. and Markov-Dependent Multistate Trials. *Advances in Applied Probability*, *41*, 292–308.

Fu, J. C., Koutras, M. V. (1994). Distribution theory of runs: a Markov chain approach. *Journal of the American Statistical Association*, *89*, 1050–1058.

Fu, J. C., Lou, W. Y. W. (2003). *Distribution Theory of Runs and Patterns and Its Applications* (1st ed.). Singapore: World Scientific.

Ivchenko, G. I. (1998). How many samples does it take to see all of the balls in an urn? *Mathematical Notes*, *64*, 49–54.

Johnson, B. C., Sellke, T. M. (2010). On the Number of i.i.d. Samples Required to Observe All of the Balls in an Urn. *Methodology and Computing in Applied Probability*, *12*, 139–154.

Klaassen, C. A. J. (1994). Dixie cups: sampling with replacement from a finite population. *Journal of Applied Probability*, *31*, 940–948.

Markoff, A.A. (1912). *Wahrscheinlichkeitsrechnung*, Berlin: Leipzig.

Newman, D. J., Shepp, L. (1960). The Double Dixie Cup Problem. *The American Mathematical Monthly*, *67*, 58–61.

Pólya, G. (1930). Eine Wahrscheinlichkeitsaufgabe in der Kundenwerbung. *Zeitschrift für Angewandte Mathematik und Mechanik*, *10*, 96–97.

Sellke, T. M. (1995). How many iid samples does it take to see all the balls in a box. *The Annal of Applied Probability*, *5*, 294–309.