J. Appl. Prob. **40**, 513–518 (2003)
Printed in Israel

THE COUPON-COLLECTOR'S PROBLEM REVISITED

ILAN ADLER,*
SHMUEL OREN * AND
SHELDON M. ROSS,* ** *University of California, Berkeley*

Abstract

Consider the classical coupon-collector's problem in which items of m distinct types arrive in sequence. An arriving item is installed in system $i \geq 1$ if i is the smallest index such that system i does not contain an item of the arrival's type. We study the expected number of items in system j at the moment when system 1 first contains an item of each type.

Keywords: Coupon-collector's problem; hyperharmonic numbers; Poissonization

AMS 2000 Subject Classification: Primary 60C05

1. Introduction

Consider the classical coupon-collector's problem with m distinct types of items. The items arrive in sequence, with the types of the successive items being independent random variables that are each equal to k with probability p_k, $\sum_{k=1}^{m} p_k = 1$. An arriving item is installed in system $i \geq 1$ if i is the smallest index such that system i does not contain an item of the arrival's type. Let U_j^m, $j \geq 2$, denote the number of unfilled types in system j when system 1 first contains an item of each type. Foata *et al.* [2] and Foata and Zeilberger [1], using nonelementary mathematics, obtained recursive formulae and generating functions for $\mathrm{E}[U_j^m]$ for the equally likely case, where $p_k = 1/m$. In Section 2 we derive, using basic probability, the recursion and a closed-form expression for $\mathrm{E}[U_j^m]$ for the equally likely case. The general case is considered in Section 3 where an exact expression and bounds for $\mathrm{E}[U_j^m]$ are determined. Comments concerning computation, as well as a simulation approach, are also presented in Section 3.

2. The equally likely case

Assume, in this section, that all $p_k = 1/m$. Furthermore, assume that the problem ends when system 1 has one item of each type, and let A_j^k denote the event that at least j type-k coupons have arrived. With $\mathbf{1}(A)$ denoting the indicator variable for the event A,

$$U_j^m = \sum_{k=1}^{m} [1 - \mathbf{1}(A_j^k)].$$

Received 10 December 2002; revision received 20 January 2003.
* Postal address: Department of Industrial Engineering and Operations Research, University of California, Berkeley, CA 94720, USA.
** Email address: smross@ieor.berkeley.edu
Research supported by the National Science Foundation Grant ECS-0224779 with the University of California.

Thus,

$$\mathrm{E}[U_j^m] = \sum_{k=1}^{m} [1 - \mathrm{P}(A_j^k)]$$
$$= m[1 - \mathrm{P}(A_j^m)]. \qquad (1)$$

Let $B_{j,i}^m$ denote the event that at least j type-m coupons arrive before the first coupon of type i arrives. Then

$$\mathrm{P}(A_j^m) = \mathrm{P}\bigg(\bigcup_{i=1}^{m-1} B_{j,i}^m\bigg)$$

and the inclusion–exclusion probability equality give (for $j \geq 2$)

$$\mathrm{P}(A_j^m) = \sum_{k=1}^{m-1} (-1)^{k+1} \sum_{i_1 < i_2 < \cdots < i_k} \mathrm{P}(B_{j,i_1}^m \cdots B_{j,i_k}^m)$$
$$= \sum_{k=1}^{m-1} (-1)^{k+1} \binom{m-1}{k} \left(\frac{1}{k+1}\right)^j.$$

Using (1), this gives the following result.

Proposition 1. *For* $j \geq 2$,

$$\mathrm{E}[U_j^m] = \sum_{i=1}^{m} \binom{m}{i} \frac{(-1)^{i+1}}{i^{j-1}}.$$

Next, using basic probability arguments, we obtain a recursive expression for $\mathrm{E}[U_j^m]$ that was first presented in [1] and [2]. Let C_j^k be the event that at least j type-k coupons have already arrived at the moment when each of the item types $1, \ldots, k-1$ has arrived. Also, let X^k be the number of types $1, \ldots, k-1$ that have not yet arrived when the first coupon of type k arrives. With $P_j^k = \mathrm{P}(C_j^k)$, we obtain that

$$P_j^k = \sum_{r=0}^{k-1} \mathrm{P}(C_j^k \mid X^k = r)\,\mathrm{P}(X^k = r)$$
$$= \frac{1}{k} \sum_{r=0}^{k-1} P_{j-1}^{r+1}$$
$$= \frac{1}{k} \sum_{r=1}^{k} P_{j-1}^{r}, \qquad (2)$$

where $P_1^k = (k-1)/k$ for $k = 1, 2, \ldots$.

Substituting $A_j^m = C_j^m$ for $j \geq 2$ into (1) gives

$$\mathrm{E}[U_j^m] = m[1 - P_j^m], \qquad j \geq 2. \qquad (3)$$

Thus, using (2) and (3), we obtain that

$$\mathrm{E}[U_2^m] = m - \sum_{r=1}^{m} \frac{r-1}{r} = \sum_{k=1}^{m} \frac{1}{k}$$

and, for $j \geq 3$,

$$\begin{aligned} \mathrm{E}[U_j^m] &= m - \sum_{k=1}^{m} P_{j-1}^k \\ &= m - \sum_{k=1}^{m}\left(1 - \frac{\mathrm{E}[U_{j-1}^k]}{k}\right) \\ &= \sum_{k=1}^{m} \frac{\mathrm{E}[U_{j-1}^k]}{k}. \end{aligned}$$

We have thus proven the following.

Proposition 2. *We have*

$$\mathrm{E}[U_2^m] = \sum_{k=1}^{m} \frac{1}{k}$$

and, for $j \geq 3$,

$$\mathrm{E}[U_j^m] = \sum_{k=1}^{m} \frac{\mathrm{E}[U_{j-1}^k]}{k}.$$

Remark 1. Equating the two expressions for $\mathrm{E}[U_j^m]$ given by Propositions 1 and 2 yields an explicit expression for the *hyperharmonic number*, which is defined in [2] by the recursive formula given in Proposition 2.

3. The general case: Poissonization

In the general case, we suppose that each item is of type k with probability p_k, $\sum_{k=1}^{m} p_k = 1$. To analyze this case, let us start by assuming that, rather than stopping when system 1 is filled, items continue coming forever. Suppose also that successive items arrive at times distributed according to a Poisson process with rate 1. Under this scenario, the arrival processes of the distinct types are independent Poisson processes, with respective rates p_k, $k = 1, \ldots, m$. Because $1 - \mathrm{P}(A_j^k)$ denotes the probability that there have been less than j type-k arrivals when system 1 becomes full, we obtain upon conditioning on the arrival time of the jth item of type k that

$$1 - \mathrm{P}(A_j^k) = \int_0^\infty p_k \mathrm{e}^{-p_k x} \frac{(p_k x)^{j-1}}{(j-1)!} \prod_{i \neq k} (1 - \mathrm{e}^{-p_i x})\, \mathrm{d}x, \qquad j \geq 2. \tag{4}$$

The expected number of unfilled slots in system j is now obtained from

$$\mathrm{E}[U_j^m] = \sum_{k=1}^{m} [1 - \mathrm{P}(A_j^k)], \qquad j \geq 2. \tag{5}$$

The following lemma will be used to obtain bounds on $\mathrm{E}[U_j^m]$.

Lemma 1. *For positive values* x_i, $\prod_{i=1}^{r}(1 - \mathrm{e}^{-x_i})$ *is a Schur concave function of* $\mathbf{y} = (y_1, \ldots, y_r)$, *where* $y_i = \ln(x_i)$.

Proof. With $y = \ln(x)$,

$$\frac{\partial}{\partial y}(1 - e^{-x}) = x e^{-x}.$$

Because $\ln(x)$ in increasing in x, by the Ostrowski condition for Schur concavity (see [3]) it suffices to show that

$$x_1 e^{-x_1}(1 - e^{-x_2}) > x_2 e^{-x_2}(1 - e^{-x_1}) \quad \text{if } x_1 < x_2.$$

But this inequality follows because $xe^{-x}/(1 - e^{-x})$ is a decreasing function of x.

Lower and upper bounds on $E[U_j^m]$, fairly tight for values of $(p_1, p_2, \dots, p_m)$ close to $(1/m, 1/m, \dots, 1/m)$, can be obtained from the inequalities

$$(1 - e^{-m_k x})^{m-1} \le \prod_{i \ne k}(1 - e^{-p_i x}) \le (1 - e^{-g_k x})^{m-1}, \tag{6}$$

where $m_k = \min_{i \ne k}\{p_i\}$ and $g_k = (\prod_{i \ne k} p_i)^{1/(m-1)}$. That is, g_k is the geometric mean of the values p_i for $i \ne k$. The second inequality of (6) follows from Lemma 1.

We obtain from (4) and (6) that

$$\begin{aligned}
1 - P(A_j^k) &\le \int_0^\infty p_k e^{-p_k x} \frac{(p_k x)^{j-1}}{(j-1)!}(1 - e^{-g_k x})^{m-1}\, dx \\
&= \sum_{r=0}^{m-1} \binom{m-1}{r}(-1)^r \int_0^\infty p_k e^{-(r g_k + p_k)x} \frac{(p_k x)^{j-1}}{(j-1)!}\, dx \\
&= \sum_{r=0}^{m-1} \binom{m-1}{r}(-1)^r \left(\frac{p_k}{r g_k + p_k}\right)^j \int_0^\infty \lambda e^{-\lambda x} \frac{(\lambda x)^{j-1}}{(j-1)!}\, dx \\
&= \sum_{r=0}^{m-1} \binom{m-1}{r}(-1)^r \left(\frac{p_k}{r g_k + p_k}\right)^j,
\end{aligned}$$

where $\lambda = r g_k + p_k$. Substituting the preceding inequality into (5) and considering both inequalities of (6) gives

$$\sum_{r=0}^{m-1} \binom{m-1}{r}(-1)^r \sum_{k=1}^{m} \left(\frac{p_k}{r m_k + p_k}\right)^j \le E[U_j^m] \le \sum_{r=0}^{m-1} \binom{m-1}{r}(-1)^r \sum_{k=1}^{m} \left(\frac{p_k}{r g_k + p_k}\right)^j. \tag{7}$$

We will now derive a second set of lower and upper bounds for $E[U_j^m]$. Let $B_{j,i}^k$ denote the event that at least j coupons of type k arrive before the first of type i arrives. Then, using the conditional expectation inequality (Proposition 3.2.3 of [5]), we obtain that

$$\begin{aligned}
P(A_j^k) &= P\left(\bigcup_{i \ne k} B_{j,i}^k\right) \\
&\ge \sum_{i \ne k} \frac{P(B_{j,i}^k)}{1 + \sum_{r \ne i,k} P(B_{j,r}^k \mid B_{j,i}^k)} &(8) \\
&= \sum_{i \ne k} \frac{(p_k/(p_k + p_i))^j}{1 + \sum_{r \ne i,k}((p_k + p_i)/(p_k + p_i + p_r))^j}, &(9)
\end{aligned}$$

where (8) follows from the conditional expectation inequality and (9) from

$$\begin{aligned}\mathrm{P}(B_{j,r}^k \mid B_{j,i}^k) &= \frac{\mathrm{P}(B_{j,r}^k B_{j,i}^k)}{\mathrm{P}(B_{j,i}^k)} \\ &= \frac{(p_k/(p_k+p_i+p_r))^j}{(p_k/(p_k+p_i))^j} \\ &= \left(\frac{p_k+p_i}{p_k+p_i+p_r}\right)^j.\end{aligned}$$

Therefore, we obtain our second upper bound for $\mathrm{E}[U_j^m] = \sum_{k=1}^m [1-\mathrm{P}(A_j^k)]$:

$$\mathrm{E}[U_j^m] \le m - \sum_{k=1}^m \sum_{i\ne k} \frac{(p_k/(p_k+p_i))^j}{1+\sum_{r\ne i,k}(p_k+p_i)^j/(p_k+p_i+p_r)^j}. \tag{10}$$

To obtain a lower bound, let X_i denote the time of the first type-i event, and let T_j^k denote the time of the jth type-k event in the Poissonization scheme (which results in T_j^k and X_i for $i \ne k$ being independent). Then, from (4),

$$1-\mathrm{P}(A_j^k) = \mathrm{E}\left[\prod_{i\ne k}(1-\mathrm{e}^{-p_i T_j^k})\right].$$

Using the well-known result that $\mathrm{E}[f(X)g(X)] \ge \mathrm{E}[f(X)]\,\mathrm{E}[g(X)]$ whenever f and g are increasing functions [4, p. 339], which easily generalizes to the product of any number of positive increasing functions, the preceding equation yields that

$$\begin{aligned}1-\mathrm{P}(A_j^k) &\ge \prod_{i\ne k}\mathrm{E}[1-\mathrm{e}^{-p_i T_j^k}] \\ &= \prod_{i\ne k}\mathrm{P}(T_j^k > X_i) \\ &= \prod_{i\ne k}[1-\mathrm{P}(T_j^k < X_i)] \\ &= \prod_{i\ne k}\left[1-\left(\frac{p_k}{p_i+p_k}\right)^j\right].\end{aligned}$$

Thus, we have the lower bound

$$\mathrm{E}[U_j^m] \ge \sum_{k=1}^m \prod_{i\ne k}\left[1-\left(\frac{p_k}{p_i+p_k}\right)^j\right]. \tag{11}$$

Remark 2. (i) Our computational experiments verify that the bounds given in (7) work well for probabilities p_i which are roughly the same, while the bounds given in (10) and (11) are tighter otherwise.

(ii) For the equal-probabilities case, the explicit expression for $\mathrm{E}[U_j^m]$ of Proposition 1 is faster to compute than the recursive expression of Proposition 2. However, for large m (say $m \ge 150$), the explicit expression (but not the recursive one) is computationally unstable.

(iii) For very large m, simulation can be employed to efficiently estimate $\mathrm{E}[U_j^m]$. The following simulation approach estimates $1 - \mathrm{P}(A_j^k)$ by a conditional expectation estimator that conditions on the arrival time of the jth item of type k; the estimator is then further improved by the use of antithetic variables.

- Generate random numbers $U_1, \ldots, U_j$;
- let $L_1 = \ln(\prod_{i=1}^{j} U_i)$ and $L_2 = \ln(\prod_{i=1}^{j}(1 - U_i))$;
- set

$$V = \frac{1}{2}\sum_{k=1}^{m}\left[\prod_{i \neq k}(1 - \mathrm{e}^{p_i L_1/p_k}) + \prod_{i \neq k}(1 - \mathrm{e}^{p_i L_2/p_k})\right].$$

The preceding should be repeated many times, with the estimator of $\mathrm{E}[U_j^m]$ being the average of the values of V obtained.

References

[1] Foata, D. and Zeilberger, D. (2002). The collector's brotherhood problem using the Newman–Shepp symbolic method. To appear in *Algebra Univ.*

[2] Foata, D., Guo-Niu, H. and Lass, B. (2001). Les nombres hyperharmonique et la fratrie du collectionneur de vignettes. *Sem. Lothar. Combinatoire* **47,** B47a (electronic).

[3] Marshall, A. W. and Olkin, I. (1979). *Inequalities: Theory of Majorization and Its Applications* (Math. Sci. Eng. **143**). Academic Press, New York.

[4] Ross, S. M. (1996). *Stochastic Processes*, 2nd edn. John Wiley, New York.

[5] Ross, S. M. (2002). *Probability Models for Computer Science*. Academic Press, New York.