J. Appl. Prob. **38,** 737–746 (2001)
Printed in Israel

THE COUPON SUBSET COLLECTION PROBLEM

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

Abstract

The coupon subset collection problem is a generalization of the classical coupon collecting problem, in that rather than collecting individual coupons we obtain, at each time point, a random subset of coupons. The problem of interest is to determine the expected number of subsets needed until each coupon is contained in at least one of these subsets. We provide bounds on this number, give efficient simulation procedures for estimating it, and then apply our results to a reliability problem.

Keywords: Coupon collecting simulation; combinatorial probability; bounds; exchangeable case; randomly chosen subsets; reliability application

AMS 2000 Subject Classification: Primary 60C05
Secondary 65C05

1. Introduction

The coupon subset collection problem utilizes a finite number of distinct types of coupons, which are obtained by purchasing coupon packages. The sets of distinct types of coupons contained in different packages are independent and identically distributed. Mathematically, let $\mathcal{S} = \{1, 2, \dots, s\}$ and let $\mathcal{S}_j,\ j = 1, \dots, m$, be subsets of $\mathcal{S}$, and suppose that each purchased package yields the subset $\mathcal{S}_j$ with probability α_j, so $\sum_{j=1}^{m} \alpha_j = 1$. Let X denote the number of subsets chosen until every element of $\mathcal{S}$ is contained in at least one of these subsets. We are interested in the mean and distribution of X.

When all subsets are of size 1 the preceding is the classical coupon collecting problem (see, for example, [2]). A special case of the coupon subsets problem is studied in [7], where it is assumed that the number of types in a randomly chosen subset has a hypergeometric distribution, and, conditional on its size being t, each of the $\binom{s}{t}$ subsets of size t are equally likely. Other related literature, as well as historical comments, can be found in [3], [4] and [7].

2. The coupon subset problem

Proposition 1. *For nonnegative random variables $X_i,\ i = 1, \dots, s$,*

$$\mathrm{P}\left(\max_i X_i > x\right) = \sum_{j=1}^{s} (-1)^{j+1} \sum_{i_1 < i_2 < \cdots < i_j} \mathrm{P}(\min(X_{i_1}, \dots, X_{i_j}) > x), \tag{1}$$

Received 12 February 2001.
* Postal address: Department of Industrial Engineering and Operations Research, University of California, Berkeley, 94720.
** Email address: ross@ieor.berkeley.edu. Research supported by the National Science Foundation Grant DMI-9901053 with the University of California.

and

$$\mathrm{E}\left[\max_i X_i\right] = \sum_{j=1}^{s}(-1)^{j+1} \sum_{i_1<i_2<\cdots<i_j} \mathrm{E}[\min(X_{i_1}, X_{i_2}, \ldots, X_{i_j})]. \tag{2}$$

Moreover, for $k = 1, 2, \ldots s$,

$$\mathrm{E}\left[\max_i X_i\right] \leq \sum_{j=1}^{k}(-1)^{j+1} \sum_{i_1<i_2<\cdots<i_j} \mathrm{E}[\min(X_{i_1}, X_{i_2}, \ldots, X_{i_j})], \tag{3}$$

if k is odd, and

$$\mathrm{E}\left[\max_i X_i\right] \geq \sum_{j=1}^{k}(-1)^{j+1} \sum_{i_1<i_2<\cdots<i_j} \mathrm{E}[\min(X_{i_1}, X_{i_2}, \ldots, X_{i_j})], \tag{4}$$

if k is even.

Proof. Letting A_i be the event that $X_i > x$, (1) is just a statement of the inclusion–exclusion probability equality. Equation (2) follows from integrating both sides of (1) from 0 to ∞. The inequalities follow upon integration of the inclusion–exclusion probability inequalities.

Remark. Equations (1) and (2) remain valid for general X_i; the inequalities (3) and (4), however, require nonnegativity.

Let X_i denote the number of subsets that must be chosen to obtain one that contains coupon i, and define

$$X = \max_i X_i.$$

Let $p(i_1, \ldots, i_j)$ denote the probability that a randomly chosen subset contains any of the type coupons $i_1, \ldots, i_j$. That is,

$$p(i_1, \ldots, i_j) = \sum_{k: S_k \cap \{i_1, \ldots, i_j\} \neq \varnothing} \alpha_k.$$

Because $\min(X_{i_1}, \ldots, X_{i_j})$ is geometric with parameter $p(i_1, \ldots, i_j)$, we obtain the following from Proposition 1.

Corollary 1. *For an integer r*

$$\mathrm{E}[X] = \sum_{j=1}^{s}(-1)^{j+1} \sum_{i_1<i_2<\cdots<i_j} \frac{1}{p(i_1, i_2, \ldots, i_j)}, \tag{5}$$

$$\mathrm{P}(X > r) = \sum_{j=1}^{s}(-1)^{j+1} \sum_{i_1<i_2<\cdots<i_j} (1 - p(i_1, i_2, \ldots, i_j))^r. \tag{6}$$

Although (5) involves a summation of 2^s terms, it can be efficiently evaluated in a few special cases which we now consider.

Note. In the following we adopt the convention that $\binom{a}{b} = 0$ whenever $a < b$.

2.1. The classical coupon collecting problem

The classical coupon collecting problem is a special case in which $|S_i| = 1$ for all $i = 1, 2, \dots, s$. Substituting $p(i_1, i_2, \dots, i_j) = p(i_1) + p(i_2) + \cdots + p(i_j)$ in (5), and using $1/p = \int_0^\infty \mathrm{e}^{-px}\,\mathrm{d}x$, gives the following well-known expression (see, for example, [2] and [3]):

$$\mathrm{E}[X] = \int_0^\infty \left[1 - \prod_{i=1}^{s}(1 - \mathrm{e}^{-p(i)x})\right]\mathrm{d}x.$$

2.2. Independent subsets

In this case, a subset is selected by independently including a type i coupon in the subset with probability $p(i)$. Consequently, the X_i are independent random variables, giving

$$\mathrm{E}[X] = \sum_{r=0}^{\infty}\left[1 - \prod_{i=1}^{s}[1 - (1 - p(i))^r]\right].$$

2.3. Exchangeable elements

Suppose that, conditional on the event that there are k coupon types in a randomly chosen subset, the set of coupon types is equally likely to be any of the $\binom{s}{k}$ subsets of size k. Let β_k be the probability that the random subset is of size k, then

$$p(i_1, i_2, \dots, i_j) = 1 - \sum_{k=1}^{s}\beta_k\frac{\binom{s-j}{k}}{\binom{s}{k}}.$$

Consequently,

$$\mathrm{E}[X] = \sum_{j=1}^{s}(-1)^{j+1}\frac{\binom{s}{j}}{1 - \sum_{k=1}^{s}\beta_k\binom{s-j}{k}/\binom{s}{k}},$$

$$\mathrm{P}(X > r) = \sum_{j=1}^{s}(-1)^{j+1}\binom{s}{j}\left(\sum_{k=1}^{s}\beta_k\frac{\binom{s-j}{k}}{\binom{s}{k}}\right)^r.$$

Note that if all the subsets are of equal size, say $\beta_m = 1$, then

$$\mathrm{E}[X] = \binom{s}{m}\sum_{j=1}^{s}(-1)^{j+1}\frac{\binom{s}{j}}{\binom{s}{m} - \binom{s-j}{m}}, \tag{7}$$

$$\mathrm{P}(X > r) = \sum_{j=1}^{s}(-1)^{j+1}\binom{s}{j}\frac{\binom{s-j}{m}^r}{\binom{s}{m}^r}, \tag{8}$$

where (7) was previously obtained in [7].

Remark. The random variables $X_1, \dots, X_s$ are clearly exchangeable under the conditions of this subsection. In fact, these will be the only cases for which they will be exchangeable. To see this, suppose they are exchangeable. Let $\mathbf{1}_j$ be the indicator for the event that $X_j = 1$, and note that $\mathbf{1}_1, \dots, \mathbf{1}_s$ are also exchangeable. Then $\mathrm{P}(\mathbf{1}_{i_1} = \cdots = \mathbf{1}_{i_k} = 1, \mathbf{1}_j = 0,\ j \neq i_1, \dots, i_k)$ depends only on k. Calling this probability P_k, and letting $\beta_k = \mathrm{P}(\sum_{i=1}^{s}\mathbf{1}_i = k)$, we have

$$P_k = \beta_k\mathrm{P}\left(\mathbf{1}_{i_1} = \cdots = \mathbf{1}_{i_k} = 1, \mathbf{1}_j = 0,\ j \neq i_1, \dots, i_k \;\middle|\; \sum_{i=1}^{s}\mathbf{1}_i = k\right).$$

Hence, the conditional probability on the right-hand side depends only on k, showing that, conditional on k coupon types, the set of coupon types is equally likely to be any of the $\binom{s}{k}$ subsets of size k.

2.4. The magic cards collecting problem

Let $M_1, \dots, M_n$ be a partition of $\mathcal{S}$ and suppose that each random subset is equally likely to be any subset of $\mathcal{S}$ that contains exactly m_k elements of M_k, for $k = 1, \dots, n$. An application of this case, which is related to the game of collecting a whole set of magic cards by purchasing sealed packages, is presented in [1]. Let $s_k = |M_k|$.

Let $Y_k = \max_{i \in M_k} X_i$. Then $X = \max_k Y_k$, implying that

$$\mathrm{P}(X > r) = 1 - \mathrm{P}(X \le r) = 1 - \prod_{k=1}^{n} [1 - \mathrm{P}(Y_k > r)].$$

Because Y_k corresponds to the time to obtain a full set in the exchangeable case having $s = s_k$, $\beta_{m_k} = 1$, we see from (8) that

$$\mathrm{P}(Y_k > r) = \sum_{j=1}^{s_k} (-1)^{j+1} \binom{s_k}{j} a(j,k)^r,$$

where

$$a(j,k) = \frac{\binom{s_k - j}{m_k}}{\binom{s_k}{m_k}}, \qquad j = 1, \dots, s_k, \ k = 1, \dots, n$$

we can now compute $\mathrm{E}[X]$ from

$$\mathrm{E}[X] = \sum_{r=0}^{\infty} \mathrm{P}(X > r) = \sum_{r=0}^{\infty} \bigg[1 - \prod_{k=1}^{n} [1 - \mathrm{P}(Y_k > r)] \bigg].$$

3. Upper bounds for the mean

As stated previously, the expression for $\mathrm{E}[X]$ as given in (5) involves 2^s terms and thus it is generally impractical for exact computations. A potentially more practical approach is to use the first few terms of (3) as upper and lower bounds for $\mathrm{E}[X]$. However, it turns out that the resulting bounds are too loose to be beneficial. As an alternative, we offer in the following upper bounds obtained by applying (3) to a simple upper bound for $\mathrm{E}[X]$ rather than to $\mathrm{E}[X]$ itself.

Again let X_i denote the number of subsets that must be chosen to obtain one that contains coupon i, and let $X = \max_i X_i$.

Proposition 2. *For any $k \ge 0$,*

$$\mathrm{E}[X] \le k + \sum_{i=1}^{s} \mathrm{E}[(X_i - k)^+],$$

$$\begin{aligned}\mathrm{E}[X] \le k &+ \sum_{i} \mathrm{E}[(X_i - k)^+] - \sum_{i<j} \mathrm{E}[\min((X_i - k)^+, (X_j - k)^+)] \\ &+ \sum_{i<j<\ell} \mathrm{E}[\min((X_i - k)^+, (X_j - k)^+, (X_\ell - k)^+)].\end{aligned}$$

Proof. Take expectations of both sides of the inequality

$$X \leq k + \max_i (X_i - k)^+$$

and then apply (3).

Corollary 2. *For any* $k = 0, 1, \ldots,$

$$\mathrm{E}[X] \leq k + \sum_{i=1}^{s} \frac{(1 - p(i))^k}{p(i)}, \tag{9}$$

$$\mathrm{E}[X] \leq k + \sum_{i} \frac{(1 - p(i))^k}{p(i)} - \sum_{i<j} \frac{(1 - p(i, j))^k}{p(i, j)} + \sum_{i<j<\ell} \frac{(1 - p(i, j, \ell))^k}{p(i, j, \ell)}. \tag{10}$$

Proof. Conditioning on whether $\min(X_{i_1}, \ldots, X_{i_j}) > k$ yields that

$$\mathrm{E}[\min((X_{i_1} - k)^+, \ldots, (X_{i_j} - k)^+)] = \frac{(1 - p(i_1, \ldots, i_j))^k}{p(i_1, \ldots, i_j)},$$

and the result follows from Proposition 2.

Let $\mathrm{UB}_1(k)$, $\mathrm{UB}_2(k)$ denote the upper bounds presented in (9) and (10) respectively. While these bounds are valid for any nonnegative integer k, we are obviously interested in making them as small as possible.

Noting that

$$\frac{\partial^2}{\partial k^2} \mathrm{UB}_1(k) = \sum_{i=1}^{s} \frac{(1 - p(i))^k (\ln(1 - p(i)))^2}{p(i)} \geq 0$$

(thus establishing that $\mathrm{UB}_1(k)$ is convex), we proceed to obtain k^*, the integer minimizer of $\mathrm{UB}_1(k)$, as follows:

- Set $U = \mathrm{UB}_1(k)$ (where k is an arbitrary nonnegative integer).
- Starting with the interval $[0, U]$, use the golden ratio search to find an interval, say u, of length less than 1, which contains the true (continuous) minimizer of $\mathrm{UB}_1(k)$.
- If u contains an integer, then this integer is k^*. Otherwise let k^* be determined by rounding down (or up, depending on which yields a lower value for $\mathrm{UB}_1(k)$) the middle point of u.

Because we have not been able to establish that $\mathrm{UB}_2(k)$ is unimodal, we cannot guarantee that the following procedure gives the the minimizing value. However, it performed well in our numerical testing (see Section 6).

- Set $k_0 = \lfloor \mathrm{UB}_1(k^*) \rfloor$.
- Evaluate $\mathrm{UB}_2(k)$ for $k = k_0, k_0 - 1, \ldots,$ as long as $\mathrm{UB}_2(k)$ decreases.
- Let $\bar{k}$ be the last value to be evaluated, then set the improved upper bound as $\mathrm{UB}_2(\bar{k} - 1)$.

4. Simulating $\mathrm{E}[X]$

Suppose we simulate the process until a complete set is obtained. In addition to the raw simulation estimator for $\mathrm{E}[X]$, we offer three other unbiased simulation estimators.

4.1. First simulation estimator

Let $S_{i_1}, \ldots, S_{i_r}$ denote the sequence of distinct subsets obtained, in their order of appearance, and let T_j denote the number of subsets collected after $S_{i_{j-1}}$ has been obtained until S_{i_j} is obtained, $j = 1, \ldots, r$. Then the first simulation estimator of $\mathrm{E}[X]$ is

$$\begin{aligned}\mathrm{EST}(1) &= \mathrm{E}[X \mid S_{i_1}, \ldots, S_{i_r}] \\ &= \sum_{j=1}^{r} \mathrm{E}[T_j \mid S_{i_1}, \ldots, S_{i_r}] \\ &= 1 + \frac{1}{1-\alpha_{i_1}} + \frac{1}{1-\alpha_{i_1}-\alpha_{i_2}} + \cdots + \frac{1}{1-\sum_{k=1}^{r-1}\alpha_{i_k}}.\end{aligned}$$

The final equality following because, given the subsets that have already appeared, the time to obtain a new subset and the identity of this subset are independent.

4.2. Second simulation estimator

Number the elements of the set S so that

$$p(1) \le p(2) \le \cdots \le p(s).$$

Let $\mathbf{1}_j$, $j = 1, \ldots, s$, be the indicator for the event that j has not been collected at the moment when all of $1, \ldots, j-1$ have been collected. Then, letting A_j denote the additional number of subsets needed at the moment when all of $1, \ldots, j-1$ have been collected until j has also been collected, we have

$$\mathrm{E}[X] = \mathrm{E}\biggl[X_1 + \sum_{j=2}^{s} A_j\biggr] = \frac{1}{p(1)} + \sum_{j=2}^{s} \frac{1}{p(j)}\mathrm{E}[\mathbf{1}_j].$$

Our second simulation estimator is

$$\mathrm{EST}(2) = \frac{1}{p(1)} + \sum_{j=2}^{s} \frac{1}{p(j)}\mathbf{1}_j.$$

4.3. Third simulation estimator

The idea used to obtain the simulation estimator EST(2) could have been used on any permutation $i_1, \ldots, i_s$ of the elements of S. That is, if $\mathbf{1}\{i_j : i_1, \ldots, i_{j-1}\}$ is the indicator for the event that i_j is first obtained after all of $i_1, \ldots, i_{j-1}$ have been obtained, then

$$\frac{1}{p(i_1)} + \sum_{j=2}^{s} \frac{1}{p(i_j)}\mathbf{1}\{i_j : i_1, \ldots, i_{j-1}\}$$

is an unbiased estimator of $\mathrm{E}[X]$. Hence, another estimator is obtained by taking the average of the preceding over all $s!$ possible permutations. That is,

$$\mathrm{EST}(3) = \frac{1}{s!}\sum_{i_1,\ldots,i_s}\biggl(\frac{1}{p(i_1)} + \sum_{j=2}^{s} \frac{1}{p(i_j)}\mathbf{1}\{i_j : i_1, \ldots, i_{j-1}\}\biggr).$$

Letting $B(i)$ denote the number of elements that are collected (strictly) earlier than i, we can rewrite this as

$$\begin{aligned}\mathrm{EST}(3) &= \sum_{i=1}^{s} \frac{1}{p(i)}\left[\frac{1}{s} + \sum_{j=1}^{s-1} \frac{j!(s-j-1)!}{s!}\binom{B(i)}{j}\right] \\ &= \sum_{i=1}^{s} \frac{1}{p(i)} \sum_{j=0}^{B(i)} \frac{j!(s-j-1)!}{s!}\binom{B(i)}{j}.\end{aligned}$$

Remark. The simulation estimators presented are modifications of ones given in [5] for coupon collecting and in [6] for determining the mean coverage time of a semi-Markov process.

5. Applications

5.1. A reliability application

Consider a coherent reliability system whose set of components is $\mathcal{T}$, where by coherent we mean that there are subsets $T_1, \dots, T_s$ of $\mathcal{T}$, none of which is a subset of the other, such that the system is said to be functional if all the components of at least one of these subsets are functional. The subsets $T_1, \dots, T_s$ are called the *minimal path sets* of the system. Suppose that the system is subject to shocks, with each shock affecting a random subset of elements of $\mathcal{T}$. That is, the shocks are independent and

$$\mathrm{P}\{\text{shock } i \text{ affects elements in } R\} = \alpha_R, \qquad R \subset \mathcal{T}, \sum_R \alpha_R = 1.$$

Any component that is affected by a shock becomes a failed component. Starting with all working components, let X denote the number of shocks until the system is failed. That is, X is the number of shocks required until at least one component of each minimal path set is affected.

By associating with each shock the subset of minimal path sets that have at least one component that is affected by that shock, the problem reduces to the model we have been considering. Namely, each shock (subset) affects a random subset of $\mathcal{S} = \{1, \dots, s\}$ and we continue to observe shocks until all elements in $\mathcal{S}$ have been affected.

5.2. A cost application

Suppose there is a cost c incurred in obtaining a new subset and that when a full set is obtained any extra type j coupons can be resold for the price c_j, $j = 1, \dots, s$. That is, if T_j is the number of type j coupons obtained, and X is, as before, the total number of subsets, then C, the total cost, is given by

$$C = cX - \sum_{j=1}^{s} c_j(T_j - 1).$$

Taking expectations gives

$$\begin{aligned}\mathrm{E}[C] &= c\mathrm{E}[X] - \sum_{j=1}^{s} c_j \mathrm{E}[T_j] + \sum_{j=1}^{s} c_j \\ &= \left[c - \sum_{j=1}^{s} c_j p(j)\right]\mathrm{E}[X] + \sum_{j=1}^{s} c_j,\end{aligned}$$

where the final equality used Wald's equation to conclude that $\mathrm{E}[T_j] = p(j)\mathrm{E}[X]$. Hence, when the expected net cost of purchasing a subset is positive (that is, when $c > \sum_{j=1}^{s} c_j p(j)$) our upper bounds on $\mathrm{E}[X]$ yield upper bounds on $\mathrm{E}[C]$.

6. Numerical examples

To gain more insight into the several upper bounds and simulation estimates offered in Sections 3 and 4, we present in this section some numerical results based on randomly generated examples. For two of the special cases for which we know how to calculate the exact values for $\mathrm{E}[X]$ (see Subsections 2.1 and 2.2), we include a few examples comparing the exact value to the various upper bounds and simulation estimates. We also include the results of a general case, where no exact value is available, and compare the upper bounds to the simulation results. In the following tables, EST(0) denotes the raw simulation estimator X.

As can be seen, the upper bounds, particularly UB_2, perform well.

The simulation results detailed in Table 1 are for classical coupon problems where the probabilities $p(j)$ chosen were roughly of equal magnitude, with the ratio of the largest to the smallest being around 2. As there tend to be many duplications in the classical problem, EST(1) performs quite well. However, because of the lack of spread of the probabilities, EST(2) is a poor estimator in these cases. (Indeed, in the case where each coupon is equally likely to be any of the s types, an easy derivation shows that $\mathrm{var}(\mathrm{EST}(2)) = s^2 \sum_{j=1}^{s-1} j/(j+1)^2$, whereas the variance of the raw simulation estimator is $\mathrm{var}(\mathrm{EST}(0)) = s^2 \sum_{j=1}^{s-1}(1 - j/s)/j^2$. The estimator EST(3), the average of all $s!$ of the EST(2)-type estimators, has the smallest variance but is computationally more involved.

The simulation results detailed in Table 2 are for the independent case. Because there are 2^s possible subsets there will almost never be any duplications, and so EST(1) is basically

TABLE 1: The classical coupons collecting problem (200 runs).

				EST(0)		EST(1)		EST(2)		EST(3)	
s	Exact	UB_1	UB_2	Av	Std	Av	Std	Av	Std	Av	Std
80	434.6	470.8	444.6	434.3	120.2	431.9	31.2	434.2	162.0	434.3	17.6
90	498.7	539.5	510.1	493.4	151.5	498.9	31.3	505.5	201.9	499.3	19.0
100	582.0	629.8	595.5	580.9	163.4	575.5	38.7	595.1	238.4	578.2	24.7
110	647.0	699.2	662.0	640.9	154.1	643.8	43.8	661.8	236.4	646.6	26.3
120	722.8	780.6	739.6	718.8	182.3	727.3	44.8	691.6	222.9	726.7	27.0

TABLE 2: Independent subsets (200 runs).

				EST(0)		EST(1)		EST(2)		EST(3)	
s	Exact	UB_1	UB_2	Av	Std	Av	Std	Av	Std	Av	Std
80	426.4	437.3	427.1	400.8	463.9	400.8	463.9	426.5	16.6	423.1	99.7
90	124.7	136.2	125.6	129.5	108.1	129.5	108.1	124.5	23.5	123.9	29.9
100	1349.5	1395.8	1350.4	1125.3	993.9	1125.3	993.9	1352.2	71.2	1344.4	268.8
110	169.1	191.9	172.0	157.4	81.6	157.4	81.6	176.0	59.8	169.0	20.4
120	75.6	85.2	77.3	75.4	43.7	75.4	43.7	77.2	25.9	76.7	10.9

TABLE 3: The general case (300 runs).

				EST(0)		EST(1)		EST(2)		EST(3)	
s	Exact	UB_1	UB_2	Av	Std	Av	Std	Av	Std	Av	Std
80		110.2	101.5	97.0	46.7	95.5	41.6	96.9	36.4	97.0	16.3
90		91.6	84.7	82.1	38.3	81.0	35.7	81.1	31.4	79.8	12.3
100		275.6	251.3	245.5	156.6	250.0	156.6	247.2	49.3	247.9	66.0
110		288.1	252.7	245.7	158.7	242.5	126.7	245.0	79.7	246.7	54.5
120		505.8	424.0	420.3	324.2	402.3	244.9	419.2	94.1	413.7	100.7

equivalent to EST(0). The probabilities $\{p(j),\ j = 1, \dots, s,\}$ used for the cases in Tables 2 and 3 had a much greater spread than in the cases considered in Table 1; consequently, both EST(2) and EST(3) tended to perform quite well.

In practice, when a very precise estimate is desired it probably pays to initially do some runs and then use the results to estimate the variance of each of the simulation estimators. Based on these variance estimates, and the computational effort needed to determine the numerical value of the estimators, a single estimator can be used for the remaining runs. If the computational effort needed to evaluate the estimators from the simulated data could be ignored, then the best estimator would be to take a weighted average of the estimators EST(1), EST(2), and EST(3). (Because EST(1) is a conditional expectation of X, nothing is gained by including $X = \text{EST}(0)$ in the weighted average.) The optimal weights can then be estimated from the simulation.

7. The number of distinct subsets

Let Y denote the mean number of *distinct* subsets that must be chosen to obtain a complete set of at least one of each type of coupon. To analyze $\mathrm{E}[Y]$, let Y_i denote the number of distinct subsets needed to obtain one that contains coupon i, and note that

$$Y = \max_i Y_i.$$

With $B(i_1, \dots, i_j) = \{r : i_k \in S_r \text{ for some } k = 1, \dots, j\}$,

$$\begin{aligned}\mathrm{E}[\min(Y_{i_1}, Y_{i_2}, \dots, Y_{i_j})] &= 1 + \sum_{r \notin B(i_1,\dots,i_j)} \mathrm{P}(S_r \text{ before any } S_t \in B(i_1, \dots, i_j)) \\ &= 1 + \sum_{r \notin B(i_1,\dots,i_j)} \frac{\alpha_r}{\alpha_r + \sum_{t \in B(i_1,\dots,i_j)} \alpha_t}.\end{aligned}$$

Substituting the preceding into (2) provides an expression for $\mathrm{E}[Y]$.

While it is rather complicated to derive the computable upper bounds to $\mathrm{E}[Y]$ analogous to those developed in Section 3, we can easily provide such bounds under the additional assumptions that all the subsets are equally likely to be selected (that is, $\alpha_i = 1/m,\ i = 1, \dots, m$).

In this case, with $n(i_1, \dots, i_j)$ equal to the number of subsets that contain any of the coupon types $i_1, \dots, i_j$,

$$\mathrm{E}[\min(Y_{i_1} - k, \dots, Y_{i_j} - k) \mid \min(Y_{i_1}, \dots, Y_{i_j}) > k] = \frac{m - k + 1}{n(i_1, \dots, i_j) + 1}$$

and

$$\mathrm{P}(\min(Y_{i_1}, \dots, Y_{i_j}) > k) = \frac{\binom{m-n(i_1,\dots,i_j)}{k}}{\binom{m}{k}}.$$

Therefore,

$$\mathrm{E}[\min((Y_{i_1} - k)^+, \dots, (Y_{i_j} - k)^+)] = \frac{m-k+1}{n(i_1, \dots, i_j)+1} \frac{\binom{m-n(i_1,\dots,i_j)}{k}}{\binom{m}{k}}. \tag{11}$$

Proposition 2 can be used, together with (11), to provide upper bounds for $\mathrm{E}[Y]$.

References

[1] Bosch, R. A. (2000). Optimal card-collecting strategies for magic: the gathering. *College Math. J.* **31,** 15–21.

[2] Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*, 3rd edn. John Wiley, Chichester.

[3] Flajolet, P., Gardy, D. and Thimonier, L. (1992). Birthday paradox, coupon collectors, caching algorithms, and self-organizing search. *Discrete Appl. Math.* **39,** 207–229.

[4] Holst, L. (1986). On birthday, collectors', occupancy and other classical urn problems. *Int. Statist. Rev.* **54,** 15–27.

[5] Peköz, E. and Ross, S. M. (1997). Estimating the mean cover time of a semi-Markov process via simulation. *Prob. Eng. Inf. Sci.* **11,** 267–271.

[6] Peköz, E. and Ross, S. M. (1999). Mean cover times for coupon collectors and star graphs. In *Applied Probability and Stochastic Processes*, eds J. G. Shanthikumar and U. Sumita. Kluwer, Boston, pp. 83–94.

[7] Stadje, W. (1990). The collector's problem with group drawings. *Adv. Appl. Prob.* **22,** 866–874.