Probability in the Engineering and Informational Sciences, **21**, 2007, 435–440. Printed in the U.S.A.
DOI: 10.1017/S026996480700006X

THE MULTIPLE SUBSET COUPON COLLECTING PROBLEM

KUANG-CHAO CHANG
Department of Statistics and Information Science
Fu Jen Catholic University, Taipei 24205
Taiwan, ROC
E-mail: stat1016@mail.fju.edu.tw

SHELDON M. ROSS
Department of Industrial Engineering and Operations Research
University of Southern California
Los Angeles, CA 90089
E-mail: smross@usc.edu

Suppose that there are n types of coupons and that each new coupon collected is type i with probability p_i. Suppose, further, that there are m subsets of coupon types and that coupons are collected until all of the types of at least one of these subsets have been collected. When these subsets have no overlap, we derive expressions for the mean and variance of the number of coupons that are needed. In the general case where the subsets can overlap, we derive the mean of the number that are needed. We also note that this number is an increasing failure rate on average random variable and we present a conjecture as to a sufficient condition for it to be an increasing failure rate random variable.

1. INTRODUCTION

As in the classical coupon collecting problem, suppose that there are n distinct types of coupon and that, independent of the types of previous coupon collected, any new coupon will be of type i with probability p_i, $\sum_{i=1}^{n} p_{i=1}$. Let $S_1, \ldots, S_m$ be sets of coupon types (i.e., each S_i is a subset of $\{1, \ldots, n\}$ and suppose that we are interested in N, the total number of coupons that have been collected when the collection first contains all of the types of at least one of the subsets $S_1, \ldots, S_m$.

The special case where the sets $S_1, \ldots, S_m$ are mutually disjoint is considered in Section 3, and we derive the mean and variance of N in this case. In Section 4 we

© 2007 Cambridge University Press 0269-9648/07 $25.00

https://doi.org/10.1017/S026996480700006X Published online by Cambridge University Press

obtain an exact expression as well as a lower bound for $E[N]$ in the general case where the subsets might overlap.

The preceding special case is of particular interest to one of the authors who, while a school child in Taiwan in the mid-1960s, often purchased packs of Princess Snow White chewing gum. Each such pack contained not only gum but also a figure from one of four well-known ancient Chinese novels: *The Romance of the Three Kingdoms, The Journey to the West, The Dream of the Red Chamber*, and *Feng Shen Bang* (a classical novel of ancient gods and heros in Chinese history). Altogether there were 100 figures from each of these stories, and the chewing gum company gave a prize to anyone who successfully collected all of the figures from any of these stories.

Before giving our results, we need some preliminaries.

2. PRELIMINARIES

To study N, we use the standard poissonization technique of assuming that the times at which coupons are collected constitute a Poisson process with rate 1. Letting X_i denote the time it takes to collect a type i coupon, it follows (see [5]) that $X_1, \ldots, X_n$ are independent, with X_i being exponentially distributed with rate p_i. For any set S of types of coupon, let

$$T(S) = \max_{j \in S} X_j$$

so that $T(S)$ denotes the time until at least one of each coupon type in S has been collected. By the independence of the X_i,

$$P(T(S) > t) = 1 - \prod_{j \in S} (1 - e^{-p_j t}). \tag{1}$$

Letting

$$T = \min_i T(S_i),$$

it follows that T is the time until the collection contains all of the types of at least one of the subsets $S_1, \ldots, S_m$.

The relation between T and N is that

$$T = \sum_{j=1}^{N} I_j,$$

where $I_1, \ldots$ are the interarrival times of the Poisson process of when coupons are collected. Because the interarrival times and the types of coupon collected are independent, it follows that T is a compound random variable, such that

$$E[T] = E[N]E[I_j] = E[N] \tag{2}$$

https://doi.org/10.1017/S026996480700006X Published online by Cambridge University Press

and

$$\mathrm{Var}(T) = \mathrm{Var}(N)(E[I_j])^2 + E[N]\,\mathrm{Var}(I_j) = \mathrm{Var}(N) + E[N], \tag{3}$$

where the preceding used that I_j is exponential with rate 1.

3. THE NONOVERLAPPING CASE

If $S_1, \ldots, S_m$ are mutually disjoint, then the random variables $T(S_i)$, $i = 1, \ldots, m$, are independent, yielding, from (1),

$$\begin{aligned} E[T] &= \int_0^\infty P(T > t)\,dt = \int_0^\infty P(\min_i T(S_i) > t)\,dt \\ &= \int_0^\infty \prod_{i=1}^m \left[1 - \prod_{j \in S_i} (1 - e^{-p_j t})\right] dt \end{aligned} \tag{4}$$

and

$$E[T^2] = 2\int_0^\infty tP(T > t)\,dt = 2\int_0^\infty t \prod_{i=1}^m \left[1 - \prod_{j \in S_i} (1 - e^{-p_j t})\right] dt \tag{5}$$

It now follows from (2) that

$$E[N] = \int_0^\infty \prod_{i=1}^m \left[1 - \prod_{j \in S_i} (1 - e^{-p_j t})\right] dt \tag{6}$$

and from (3) that

$$\mathrm{Var}(N) = 2\int_0^\infty t \prod_{i=1}^m \left[1 - \prod_{j \in S_i} (1 - e^{-p_j t})\right] dt - E[N](1 + E[N]). \tag{7}$$

The numerical computation of $E[N]$ and $\mathrm{Var}(N)$ through the use of Eqs. (6) and (7) can be involved. For some thoughts on the computation of these quantities in the classical coupon collecting problem, see [3].

4. THE GENERAL CASE

When the sets S_i are not mutually disjoint, we can derive an expression for $E[N]$ by using the min-max identity, which states that for any real numbers $x_1, \ldots x_n$,

$$\begin{aligned} \max x_i = &\sum_i x_i - \sum_{i<j} \min(x_i, x_j) + \sum_{i<j<k} \min(x_i, x_j, x_k) - \cdots \\ &+ (-1)^{n+1} \min(x_1, \ldots, x_n). \end{aligned}$$

(For a probabilistic proof of this identity, see [4].) Applying this identity to the

https://doi.org/10.1017/S026996480700006X Published online by Cambridge University Press

values $-x_i$, $i = 1, \ldots, n$, yields the equivalent identity

$$\min x_i = \sum_i x_i - \sum_{i<j} \max(x_i, x_j) + \sum_{i<j<k} \max(x_i, x_j, x_k) - \cdots + (-1)^{n+1} \max(x_1, \ldots, x_n).$$

Applying this identity to the values $T(S_i)$, $i = 1, \ldots, m$, yields

$$T = \sum_i T(S_i) - \sum_{i<j} T(S_i \cup S_j) + \sum_{i<j<k} T(S_i \cup S_j \cup S_k) - \cdots + (-1)^{n+1} T\left(\bigcup_{i=1}^{n} S_i\right).$$

Consequently,

$$E[T] = \sum_i E[T(S_i)] - \sum_{i<j} E[T(S_i \cup S_j)] + \sum_{i<j<k} E[T(S_i \cup S_j \cup S_k)] - \cdots + (-1)^{n+1} E\left[T\left(\bigcup_{i=1}^{n} S_i\right)\right]. \tag{8}$$

Using that, for any subset S,

$$E[T(S)] = \int_0^\infty \left[1 - \prod_{j \in S} (1 - e^{-p_j t})\right] dt,$$

we can use Eq. (8) to obtain an expression for $E[T] = E[N]$, which can be numerically determined when m is not too large.

A bound on $E[N]$ can be obtained by noting that because $T(S_1), \ldots, T(S_m)$ are all increasing functions of the independent random variables $X_1, \ldots, X_n$, it follows that they are associated (see [1]). However, this implies that

$$P(\min T(S_i) > t) \geq \prod_i P(T(S_i) > t).$$

Thus, T is stochastically larger than the minimum of independent random variables $T_1, \ldots, T_m$, where each T_i has the same distribution as $T(S_i)$, implying that

$$E[N] = E[T] \geq E[\min(T_i)] = \int_0^\infty \prod_{i=1}^{m} \left[1 - \prod_{j \in S_i} (1 - e^{-p_j t})\right] dt.$$

Remarks:

1. Let $r(n)$ be the failure rate function of N, defined by

$$r(n) = P(N = n | N > n - 1), \qquad n \geq 1.$$

https://doi.org/10.1017/S026996480700006X Published online by Cambridge University Press

It follows from [2] that N is an increasing failure rate on average (IFRA) random variable; that is, $\sum_{i=1}^{n} r(i)/n$ is a nondecreasing function of n for $n \geq 1$.

2. We conjecture that if $p_i = 1/n$ for all i and the sets $S_1, \ldots, S_m$ are mutually disjoint, then N has an increasing failure rate function. Our conjecture is based on the fact that if the times at which coupons are collected are distributed according to a Poisson process, then T is a continuous increasing failure rate (IFR) random variable. This result follows from the identity

$$T = \min_i T(S_i)$$

by noting that $T(S_i) = \max_j X_j$ is the maximum of a set of independent and identically distributed IFR random variables and so is IFR and then using the result that the minimum of independent IFR random variables is also IFR (see [1] for proofs).

3. For a counterexample showing that N need not be IFR when the S_i are nonoverlapping, consider the classical coupon collectors problem with $n = 2$. If $p_1 = p$,

$$r(2) = 2p(1-p).$$

However, for $n \geq 2$,

$$r(n+1) = (1-p)\frac{p^n}{p^n + (1-p)^n} + p\frac{(1-p)^n}{p^n + (1-p)^n} \downarrow \min(p, 1-p).$$

Thus, when $0 < p \neq 1/2$, the failure rate function $r(n)$ is strictly decreasing when $n > 2$. (This does not contradict IFRA because $r(1) = 0$.)

4. For a counterexample showing that N need not be IFR when the S_i have overlaps, even when all $p_i = 1/n$, suppose $n = 3$, $m = 2$, $S_1 = \{1, 2\}$, and $S_2 = \{1, 3\}$. With $p_i = 1/3$,

$$r(2) = 2(1/3)(2/3) = 4/9,$$

whereas

$$\lim_{n\to\infty} r(n) = 1/3,$$

which follows because if neither subset has occurred by time n, then either each of the first n coupons was type 1 or each of them was not a type 1; the latter possibility dominates as n grows.

5. In the case of general p_i, if the problem would end when for some i, there have been n_i, coupons collected that are of a type in S_i (rather than requiring that all the coupon types of S_i appear at least once), then the number of coupons needed would be IFR when the S_i are nonoverlapping. This result was proven in [6].

https://doi.org/10.1017/S026996480700006X Published online by Cambridge University Press

References

1. Barlow, R.E. & Proschan, F. (1975). *Statistical theory of reliability and life testing*. New York: Holt.
2. Brown, M. & Chaganty, N. (1983). On the first passage time distribution for a class of Markov chains. *The Annals of Probability* 11(4): 1000–1008.
3. Liu, J.-F. & Chang, K.-C. (2000). A note on multiple inverse sampling. *Journal of Statistical Planning and Inference* 87: 347–352.
4. Ross, S.M. (2006). *A first course in probability*, 7th ed. Englewood Cliffs, NJ: Prentice-Hall.
5. Ross, S.M. (2007). *Introduction to probability models*, 9th ed. San Diego: Academic.
6. Ross, S.M., Shahshahani, M., & Weiss, G. (1980). On the duration of the problem of the points. *Journal of the American Statistical Association* 75(371): 663–666.

https://doi.org/10.1017/S026996480700006X Published online by Cambridge University Press