Probability in the Engineering and Informational Sciences, **22**, 2008, 221–229. Printed in the U.S.A.
doi: 10.1017/S0269964808000132

COUPON COLLECTING

MARK BROWN
Department of Mathematics
City College
CUNY
New York, NY
E-mail: cybergaf@aol.com

EROL A. PEKÖZ
Department of Operations and Technology Management
Boston University Boston, MA 02215
E-mail: pekoz@bu.edu

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

We consider the classical coupon collector's problem in which each new coupon collected is type i with probability p_i; $\sum_{i=1}^{n} p_i=1$. We derive some formulas concerning N, the number of coupons needed to have a complete set of at least one of each type, that are computationally useful when n is not too large. We also present efficient simulation procedures for determining $P(N > k)$, as well as analytic bounds for this probability.

1. INTRODUCTION

Consider the coupon collector's problem in which there are n types of coupons and each new one collected is, independently of the past, a type j coupon with probability

Mark Brown's research was supported by the Office of Industry Relations of the National Security Agency.

 0269-9648/08 $25.00

p_j; $\sum_{j=1}^{n} p_j = 1$. Let N denote the minimum number of coupons one must collect to obtain a complete set of at least one of each type. Also, let L denote the last type to be collected. In Section 2 we derive some formulas, some of which are well known, concerning the distribution of N. These formulas typically involve sums of 2^n terms and so are only of practical use when n is moderately small. We then consider simulation techniques for when n is large. In Section 3 we present some efficient simulation techniques for estimating tail probabilities $P(N > k)$. Analytic bounds for $P(N > k)$ are also given in Sections 2 and 3. In Section 4 we propose a Markov chain Monte Carlo method for estimating the conditional tail distribution given the value of L. Many authors have studied various aspects of the coupon collector's problem. See [1] and the references cited there for additional work in this area.

2. SOME DERIVATIONS

Let B_j denote the set of all $\binom{n}{j}$ subsets of size j of $\{1, \ldots, n\}$. Also, for $\alpha \subset \{1, \ldots, n\}$, let

$$p_\alpha = \sum_{j \in \alpha} p_j, \qquad q_\alpha = 1 - p_\alpha.$$

PROPOSITION 1: *Let N be the number of coupons needed for a complete set, let L the last type to be collected before a complete set is obtained, and let X_j be the number of type j coupons among the N collected.*

(a) $P(N > k) = \sum_{j=1}^{n} (-1)^{j-1} \sum_{\alpha \in B_j} q_\alpha^k$;

(b) $P(N = k) = \sum_{j=1}^{n} (-1)^{j-1} \sum_{\alpha \in B_j} q_\alpha^{k-1} p_\alpha$;

(c) $E[s^N] = \sum_{j=1}^{n} (-1)^{j-1} \sum_{\alpha \in B_j} \frac{p_\alpha s}{1 - q_\alpha s}, \qquad s < \min(1/q_i)$;

(d) $E[N(N-1)\cdots(N-r+1)] = r! \sum_{j=1}^{n} (-1)^{j-1} \sum_{\alpha \in B_j} \frac{q_\alpha^{r-1}}{p_\alpha^r}$;

(e) $P(L = j) = \int_0^\infty p_j e^{-p_j x} \prod_{r \neq j} (1 - e^{-p_r x})\, dx$;

(f) $P(N = k, L = j) = p_j \sum_{\alpha \subset \{1, \ldots, j-1, j+1, \ldots, n\}} (-1)^{|\alpha|-1} [q_j^{k-1} - (q_j - p_\alpha)^{k-1}]$;

(g) $P(X_i = k) = \sum_{j \neq i} \int_0^\infty p_j e^{-p_j x} e^{-p_i x} \frac{(p_i x)^k}{k!} \prod_{r \neq i, j} (1 - e^{-p_r x})\, dx, \qquad k > 1$;

(h) $P(X_i = 1) = P(L = i) + \sum_{j \neq i} \int_0^\infty p_j e^{-p_j x} e^{-p_i x} p_i x \prod_{r \neq i, j} (1 - e^{-p_r x})\, dx$.

PROOF: Part (a) follows from the inclusion–exclusion theorem upon setting A_i equal to the event that there are no type i coupons among the first k collected and using

https://doi.org/10.1017/S0269964808000132 Published online by Cambridge University Press

$P(N > k) = P(\cup_i A_i)$. Part (b) follows from (a) upon using that $P(N = k) = P(N > k-1) - P(N > k)$. Part (c) follows from

$$\begin{aligned} E[s^N] &= \sum_{k=1}^{\infty} s^k P(N = k) \\ &= \sum_{j=1}^{n} (-1)^{j-1} \sum_{\alpha \in B_j} \sum_{k=1}^{\infty} s^k q_\alpha^{k-1} p_\alpha \\ &= \sum_{j=1}^{n} (-1)^{j-1} \sum_{\alpha \in B_j} \frac{p_\alpha s}{1 - q_\alpha s}, \qquad s < \min(1/q_i). \end{aligned}$$

To prove (d), use (c) along with the identity

$$E[N(N-1)\cdots(N-r+1)] = \frac{d^r}{ds^r} E[s^N]\big|_{s=1}.$$

To prove (e), we make use of the standard Poissonization trick that supposes that coupons are collected at times distributed according to a Poisson process with rate 1, which results in the collection times of type i coupons being independent Poisson processes with respective rates p_i, $i = 1,\ldots n$. Thus, if $T_{\rm i}$ denotes the first time a type i is collected, then $T_{\rm i}$, $i = 1, \ldots n$, are independent exponentials with rates $p_{\rm i}$, yielding

$$\begin{aligned} P(L = j) &= P(T_j = \max_i T_i) \\ &= \int_0^\infty p_j e^{-p_j x} \prod_{r \neq j} (1 - e^{-p_r x})\, dx. \end{aligned}$$

For (f), use that

$$P(N = k, L = j) = q_j^{k-1} p_j P(N_{-j} \leq k-1),$$

where N_{-j} is the number needed to collect a full set when each new one is one of the types i, $i \neq j$, with probability p_j/q_i. Now, use part (b).

For (g) and (h), use that for $i \neq j$,

$$P(X_i = k, L = j) = \sum_{j \neq i} \int_0^\infty p_j e^{-p_j x} e^{-p_i x} \frac{(p_i x)^k}{k!} \prod_{r \neq i,j} (1 - e^{-p_r x})\, dx.$$

https://doi.org/10.1017/S0269964808000132 Published online by Cambridge University Press

Then use

$$P(X_i = k) = \sum_{j \neq i} P(X_i = k.L = j), \qquad k > 1,$$

$$P(X_i = 1) = P(L = i) + \sum_{j \neq i} P(X_i = 1, L = j).$$

■

PROPOSITION 2: *With* $\lambda = \sum_i q_i^k$,

$$\max\left(\max_i q_i^k,\ \lambda - \sum_{i<j} q_{\{i,j\}}^k,\ \frac{\lambda^2}{\lambda + \sum_j \sum_{i \neq j} q_{\{i,j\}}^k} \right) \leq P(N > k) \leq \lambda$$

PROOF: The first term on the left-hand side inequality follows because the probability of a union is at least as large as the probability of any event of the union, the second is from the inclusion–exclusion inequalities, and the third is the conditional expectation inequality (see [3]). The right-hand side is the first inclusion–exclusion inequality (e.g., Boole's inequality).

R*emark 3*: *Because the events* A_i *and* A_j, $j \neq i$, *are negatively correlated, and so* $q_{\{i,j\}} \leq q_i\, q_j$, *it is easy to see, when* $\lambda = \sum_i q_i^k < 1$, *that*

$$\lambda - \sum_{i<j} q_{\{i,j\}}^k \geq \frac{\lambda^2}{\lambda + \sum_j \sum_{i \neq j} q_{\{i,j\}}^k}.$$

Additional bounds on $P(N > k)$ *are given in the next section.* ■

3. SIMULATION ESTIMATION OF $P(N > k)$

To efficiently use simulation to estimate $P(N > k)$, imagine that coupons are collected at times distributed according to a Poisson process with rate 1. Start by simulating T_i, $i = 1, \ldots, n$, where T_i is exponential with rate p_i and represents the time when coupon i is first collected. Now, order them so that

$$T_{I_1} < T_{I_2} < \cdots < T_{I_n}.$$

We next present our first estimator.

https://doi.org/10.1017/S0269964808000132 Published online by Cambridge University Press

PROPOSITION 4: *The estimator*

$$\mathrm{EST}_1 = P(N > k|I_1, \ldots, I_n) = \sum_{i=1}^{n-1} (1 - a_i)^{k-1} \prod_{j \neq i} \frac{a_j}{a_j - a_i},$$

where $a_j = q_{\{I_1, \ldots I_j\}}$, $j = 1, \ldots, n-1$, *is unbiased for* $P(N > k)$.

PROOF: To evaluate this conditional probability, note first that conditional on $I_1, \ldots, I_n$, N is distributed as the sum of 1 plus $n-1$ independent geometric random variables with respective parameters $q_{\{I_1, \ldots, I_j\}}$, $j = 1, \ldots, n-1$. Now use the following proposition, which can be found in Diaconis and Fill [2]. (For a simple proof of this result, Ross and Peköz [3].)

Proposition *5:* If $X_1, \ldots, X_r$ *are independent geometric random variables with parameters* $a_1, \ldots, a_r$, *where* $a_i \neq a_j$ *if* $i \neq j$, *then, for* $k \geq r-1$,

$$P(X_1 + \cdots + X_r > k) = \sum_{i=1}^{r} (1 - a_i)^k \prod_{j \neq i} \frac{a_j}{a_j - a_i}.$$

The preceding also yield additional analytic bounds on $P(N > k)$.

COROLLARY 6: *Suppose* $p_1 \leq p_2 \leq \ldots \leq p_n$. *Then*

$$\sum_{i=1}^{n-1} (1 - c_i)^{k-1} \prod_{j \neq i} \frac{c_j}{c_j - c_i} \leq P(N > k) \leq \sum_{i=1}^{n-1} (1 - d_i)^{k-1} \prod_{j \neq i} \frac{d_j}{d_j - d_i},$$

where $c_j = q_{1,2,\ldots,j}$, $d_j = q_{n,\, n-1,\ldots,\, n-j+1}$, and $j = 1, \ldots, n-1$.

PROOF: Using the assumed monotonicity of the p_j, it follows from the proof of Proposition 4 that N is stochastically larger than 1 plus the sum of $n-1$ independent geometrics with parameters $q_{1,\ldots,j}$, $j = 1, \ldots, n-1$, and is stochastically smaller than 1 plus the sum of $n-1$ independent geometrics with parameters $q_{n,\ldots,n-j+1}$, $j = 1, \ldots, n-1$.

Next is our second estimator. ■

PROPOSITION 7: *The estimator*

$$\mathrm{EST}_2 = P(N > k|T_1, \ldots, T_n) = 1 - \sum_{i=0}^{k-n} e^{-\lambda} \lambda^i / i!,$$

https://doi.org/10.1017/S0269964808000132 Published online by Cambridge University Press

where

$$\lambda = \sum_{j=1}^{n} p_{I_j}(T_{I_n} - T_{I_j})$$

is unbiased for $P(N > k)$.

PROOF: Note that N, conditional on $T_1, \ldots, T_n$, is distributed as n plus a Poisson random variable with mean $\sum_{j=1}^{n} p_{I_j}\,(T_{I_n} - T_{I_j})$. The idea being that the first type I_j coupon was collected at time T_{I_j} and so the additional number collected until time T_{I_n} is Poisson with mean $p_{I_j}\,(T_{I_n} - T_{I_j})$. ■

Remark 8: *Although the conditional variance formula implies that our second estimator has a larger variance than does our first estimator its computation—requiring only a Poisson tail probability—is simpler than the computation of the tail probability of the convolution of geometrics that is required by the first estimator.*

REMARK 9: *The preceding gives a very efficient way of simulating* N_i. i T_i, *order them, then generate a Poisson random variable* P *with mean* $\sum_{j=1}^{n} p_{I_j}\,(T_{I_n} - T_{I_j})$, *and set* $N = n + P$.

EXAMPLE 10: Suppose $p_i = i/55$, $i = 1, \ldots, 10$, *and that we want to estimate* $P(N > 200)$. *Then based on* 10^5 *simulation runs for both, the first estimator produced a sample mean of* 0.0260 *with a sample variance of* 0.0006, *whereas the second estimator produced a sample mean of* 0.0262 *with a sample variance of* 0.022. *In addition, the running time of the first method was not significantly greater than that of the second. Thus, for these input, the second estimator is only marginally better than the raw simulation estimator whose variance is approximately* $0.026(0.974) \approx 0.025$.

Both of the preceding simulation approaches yield estimates of $P(N > k)$ for every value of k. However, if we only want to evaluate $P(N > k)$ for a specified value of k, then we have another competitive approach when k is such that $P(N > k)$ is small. It is based on the following identity. (For a proof, see [3].)

PROPOSITION 11: *For events* $A_1, \ldots, A_n$, *let* $W = \sum_{i=1}^{n} I\{A_i\}$ and set $\lambda = E[W]$. *Then*

$$P(W > 0) = \lambda E\left[\frac{1}{W} | A_I\right],$$

where I *is independent of the events* $A_1, \ldots, A_n$ *and is equally likely to be any of the*

https://doi.org/10.1017/S0269964808000132 Published online by Cambridge University Press

values $1, \ldots, n$.

Next is the estimator.

PROPOSITION 12: *Let the random variable X have mass function*

$$P(X = j) = \frac{q_j^k}{\sum_{j=1}^n q_j^k}, \qquad j = 1, \ldots, n,$$

and let $(N_1, \ldots, N_n)|X{=}j$ *be multinomial with k trials and n type outcomes with type probability 0 for* N_j and p_i/q_j, $i \neq j$ *for* N_i. *Then the estimator*

$$\text{EST}_3 = \sum_{i=1}^n q_i^k / \sum_{i=1}^n I\{N_i = 0\}$$

is unbiased for $P(N > k)$.

PROOF: Let N_i represent the number of type i coupons among the first k selected, and let $A_i{=}\{N_i{=}0\}$. Thus, $P(N > k) = P(W > 0)$, where $W{=}\sum_{i=1}^n I\{A_i\}$. Then note that if I is equally likely to be any of the values $1, \ldots, n$, we have

$$P(I = j|A_I) = \frac{P(A_j)}{\sum_i P(A_i)} = \frac{q_j^k}{\sum_i q_i^k}.$$

Then apply the preceding proposition. ■

Remarks

- Because EST_3 is largest when $W = \sum_{i=1}^n I\{N_i = 0\} = 1$, it is always between zero and $\lambda = \sum_{i=1}^n q_i^k$ and so should have small variance when λ is small (see Example 13).
- To generate N_i, $i \neq j$, conditional on $N_j = 0$, generate a multinomial with k trials and $n-1$ type outcomes, with type probabilities p_i/q_j, $i \neq j$. (One way is to generate them sequentially, using that the conditional distribution at each step is binomial.) Then set W equal to 1 plus the number of components of the generated multinomial that are equal to zero.
- EST_3 can be improved by adding a stratified sampling component; that is, if you plan to do m simulation runs, rather than generate X for each run, just arbitrarily set $X{=}j$ in $mP(X = j)$ of the runs. This will always result in an estimator with smaller variance.
- The conditional expectation inequality used in Proposition 2 is obtained by applying Jensen's inequality to the result of Proposition 11. This yields

$$P(W > 0) \geq \frac{\lambda}{E[W|A_I]}.$$

https://doi.org/10.1017/S0269964808000132 Published online by Cambridge University Press

Using that $W|A_i \leq_{\text{st}} 1 + W$, the preceding—along with Boole's inequality—implies that

$$\frac{\lambda}{1+\lambda} \leq P(N > k) \leq \lambda.$$

Example 13: Suppose $p_i = i/55$, $i = 1, \ldots, 10$, *and that we want to estimate* $P(N > k)$, $k = 50, 100, 150, 200$. *Below are the data showing the mean, the variance of the raw estimator, and the variance of estimator* 3.

k	50	100	150	200
$P(N > k)$	0.54	0.18	0.07	0.03
Var raw	0.25	0.15	0.06	0.03
Var EST_3	0.026	0.00033	0.000009	0.00000016

Thus, for these inputs, the third estimator is much better than the raw simulation estimator and the other two estimators.

4. USING SIMULATION TO ESTIMATE $P(N > k|L = j)$

Suppose now that we want to estimate $P(N > k|L = j)$. Again, let T_i, $i = 1, \ldots, n$, be independent exponentials with rates p_i and again define I_n so that

$$T_{I_1} < T_{I_2} < \cdots < T_{I_n}.$$

Given the ordering of the T_i, we can utilize either of the first two methods of the preceding section to get unbiased estimates. This is summarized by the following proposition.

Proposition 14:

$$P(N > k|L = j) = E[\text{EST}_1|T_j = \max_i T_i] = E[\text{EST}_2|T_j = \max_i T_i].$$

Thus, we need to be able to generate the T_i conditional on the event that $T_j = \max_i T_i$. To do this, we recommend a Gibb's sampler approach, leading to the following procedure.

1. Start with arbitrary positive values of $T_1, \ldots, T_n$, subject to the condition that the value assigned to T_j is the largest.
2. Let J be equally likely to be any of the values $1, \ldots, n$.
3. If $J = k \neq j$, generate an exponential random variable with rate p_k conditioned to be less than t_j and let it be the new value of T_k.

https://doi.org/10.1017/S0269964808000132 Published online by Cambridge University Press

4. If $J = j$, generate an exponential with rate p_j, add it to $\max_{i \neq j} T_i$, and take that as the new value of T_j.
5. Use these values of T_i to obtain an estimate of EST_1 or EST_2.
6. Return to Step 2.

References

1. Adler, I., Oren, S., & Ross, S.M. (2003). The coupon-collector's problem revisited. *Journal of Applied Probability.* 40(2): 513–518.
2. Diaconis, P. & Fill, J.A. (1990). Strong stationary times via a new form of duality. *Annals of Probability* 18(4): 1483–1522.
3. Ross, S. & Peköz, E.A. (2007). *A second course in probability* (probabilitybookstore.com.)

https://doi.org/10.1017/S0269964808000132 Published online by Cambridge University Press