*Journal of Industrial and Systems Engineering*
Vol. 1, No. 1, pp 1-7
Spring 2007

JISE

# The Hypergeometric Coupon Collection Problem and its Dual

**Sheldon M. Ross**

Epstein Department of Industrial and Systems Engineering , University of Southern California,
Los Angeles, CA, USA
*(smross@usc.edu)*

**ABSTRACT**

Suppose an urn contains *M* balls, of different types, which are removed from the urn in a uniform random manner. In the hypergeometric coupon collection problem, we are interested in the set of balls that have been removed at the moment when at least one ball of each type has been removed. In its dual, we are interested in the set of removed balls at the first moment that this set contains all of the balls of at least one type.

## 1. INTRODUCTION AND SUMMARY

Consider an urn that contains *M* balls, of which $m_i$ are of type *i*, for $i = 1,...,n, M = \sum_{i=1}^{n} m_i$ . The hypergeometric coupon collecting problem arises when balls are removed from the urn in uniform random manner until at least one ball of each type has been removed. Let $N_i$ denote the number of type *i* balls that are removed in the coupon collecting problem, and let $N = \sum_{i=1}^{n} N_i$ denote the total number of removed balls. In section 2 we derive the joint probability mass function of $N_1,...,N_n$ as well as the marginal mass functions and the means and variances of the $N_i$. We also give formulas for the factorial moments of *N*. The tail distribution $P(N>r)$ is explicitly obtained when all $m_i$ are equal, and upper and lower bounds are provided in the general case. We also show that $P(N>r)$ is a Schur convex function of the parameters $m_1,...,m_n$.

The dual of the hypergeometric coupon collecting problem, which stops removing balls at the moment when all the balls of at least one type have been removed, is considered in Section 3.

## 2. RESULTS FOR THE HYPERGEOMETRIC COUPON COLLECTION PROBLEM

We start with the joint mass function of $N_1,...,N_n$.

**Proposition 1** With $r = \sum_j i_j$ ,

$$P(N_1 = i_1, N_2 = i_2, ..., N_n = i_n) = \sum_{j:i_j=1} \frac{\prod_{k \neq j} \binom{m_k}{i_k}}{\binom{M}{r-1}} \frac{m_j}{M-r+1}$$

**Proof:** With $L$ being the last of the $n$ types to have one of its balls removed, we have

$$P(N_1 = i_1, N_2 = i_2, \ldots, N_n = i_n) = \sum_{j:i_j=1} P(N_1 = i_1, N_2 = i_2, \ldots, N_n = i_n, L = j)$$

$$= \sum_{j:i_j=1} \frac{\prod_{k\neq j}\binom{m_k}{i_k}}{\binom{M}{r-1}} \frac{m_j}{M-r+1} \qquad \blacksquare$$

To obtain the marginal mass function of $N_k$ we will find it useful to consider a continuous time model in which each of the balls is removed at random time that is uniformly distributed on (0, 1), with these $M$ times being independent. Clearly, the order in which the balls are removed in this continuous time model is probabilistically the same as in the original model.

**Proposition 2**

$$P(N_k = 1) = \int_0^1 m_k (1-t)^{m_k - 1} \prod_{r\neq k} [1-(1-t)^{m_r}]dt + \sum_{j\neq k} \int_0^1 m_j (1-t)^{m_j - 1} m_k t (1-t)^{m_k - 1} \prod_{r\neq j,k} [1-(1-t)^{m_r}]dt$$

$$P(N_k = i) = \sum_{j\neq k} \int_0^\infty m_j (1-t)^{m_j - 1} \binom{m_k}{i} t^i (1-t)^{m_k - i} \prod_{r\neq j,k} [1-(1-t)^{m_r}]dt, \qquad i > 1$$

**Proof:** Let $T_j$ denote the time of the first removal of type $j$ ball. Then, with $L$ designating the last type ball removed

$$\begin{aligned} P(N_k = i) = \quad & P(N_k = i, L = k) + \sum_{j\neq k} P(N_k = i, L = j) \\ = \quad & P(N_k = i, L = k) + \sum_{j\neq k} \int_0^1 P(N_k = i, L = j | T_j = t) m_j (1-t)^{m_j - 1} dt \\ = \quad & P(N_k = i, L = k) + \sum_{j\neq k} \int_0^1 m_j (1-t)^{m_j - 1} \binom{m_k}{i} t^i (1-t)^{m_k - i} \prod_{r\neq j,k} [1-(1-t)^{m_r}]dt \end{aligned}$$

where

$$P(N_k = i, L = k) = \begin{cases} \int_0^1 m_k (1-t)^{m_k - 1} \prod_{r\neq k} [1-(1-t)^{m_r}]dt, & \text{if } i = 1 \\ 0, & \text{if } i > 1 \end{cases}$$

and the proof is complete. ■

**Remark:** The preceding result can be used to obtain an expression for $\sum_k P(N_k = i)$, the expected number of types to have exactly $i$ balls removed. Expressions for these quantities in the classical coupon collecting problem - where each new coupon is, independently of the past, of type $i$ with probability $p_i$, $i$ =1,...,$n$, - were given in Adler et al. (2003).
We now obtain the mean and variance of $N_k$.

**Proposition 3** With

$$p_k = \int_0^1 (1-t)^{m_k-1} \prod_{j\neq k} (1-(1-t)^{m_j})dt + \sum_{j\neq k} \int_0^1 m_j (1-t)^{m_j-1} t \prod_{r\neq k,j} (1-(1-t)^{m_r})dt,$$

$$E[N_k] = m_k p_k$$

$$Var(N_k) = m_k p_k (1-p_k) + \frac{m_k(m_k-1)}{2}[\sum_{j\neq k} \int_0^1 m_j (1-t)^{m_j-1} t^2 \prod_{r\neq k,j} (1-(1-t)^{m_r})dt - p_k^2]$$

**Proof:** Number the $m_k$ type $k$ balls, and let $R_i$ denote the event that type $k$ ball number $i$ is in the final set, $i$ =1,...,$m_k$. Then,

$$\begin{aligned} P(R_i) &= P(R_i, L=k) + \sum_{j\neq k} P(R_i, L=j) \\ &= \int_0^1 (1-t)^{m_k-1} \prod_{j\neq k} (1-(1-t)^{m_j})dt + \sum_{j\neq k} \int_0^1 m_j (1-t)^{m_j-1} t \prod_{r\neq k,j} (1-(1-t)^{m_r})dt \end{aligned}$$

Also, if $m_k \geq 2$, then for $i \neq s$

$$\begin{aligned} P(R_i R_s) &= \sum_{j\neq k} P(R_i R_s, L=j) \\ &= \sum_{j\neq k} \int_0^1 m_j (1-t)^{m_j-1} t^2 \prod_{r\neq k,j} (1-(1-t)^{m_r})dt \end{aligned}$$

the result follows since $N_k = \sum_{i=1}^{m_k} I_{R_i}$. ■

The next proposition yields a formula for factorial moments of $N = \sum_k N_k$, the total number of balls one needs to remove to obtain a complete set.

**Proposition 4**

$$E[N(N+1)...(N+r-1)] = \frac{(M+r)!}{M!} r \int_0^1 [1 - \prod_{i=1}^{n} (1-p^{m_i})](1-p)^{r-1} dp$$

**Proof:** Consider a system with *M* components in which each component is either working or failed, and suppose there is a monotone structure function that defines when the system works as a function of which components are working. Suppose that the *M* initially working components fail in a random, uniformly distributed order, and let *N* equal to the number of components that need fail to cause the system to fail. It was shown by Ross et al. (1980) that

$$E[N(N+1)...(N+r-1)] = \frac{(M+r)!}{M!} r \int_0^1 r(p)(1-p)^{r-1} dp$$

where $r(p)$ is the probability that the system works when each component independently works with probability $p$.

For a system of $M$ components, partitioned into $n$ disjoint subsets of sizes $m_1,...,m_n$, that is said to work if all of the components of at least one subset work

$$r(p) = 1 - \prod_{i=1}^{n} (1 - p^{m_i}) dp$$

and the result follows. ■

Our next result shows that the distribution function of $N$ is a Schur function of the parameters $(m_1,...,m_n)$.

**Proposition 5** $P$ (N>$r$) is a Schur convex function of $(m_1,...,m_n)$.

**Proof:** Let $N(m_1,...,m_n)$ be the number of balls removed to obtain at least one of each type. It suffices to show, for $m_1 > m_2$, that

$$P(N(m_1, m_2, m_3, ..., m_n) > r) \geq P(N(m_1 - 1, m_2 + 1, m_3, ..., m_n) > r)$$

Let $I(m_1,...,m_n)$ be the indicator variable of the event that the set of the first balls removed contains at least one of each of the types 3,...,$n$; and let $R(m_1,...,m_n)$ denote the total number of balls of types 3,...,$n$, that are among the first $r$ balls removed. Then

$$\begin{aligned} P(N(m_1,...,m_n) > r) = & \sum_{i=0}^{1} \sum_{j} P(N(m_1,....,m_n) > r | I(m_1,...,m_n) = i, R(m_1,...,m_n) = j) \\ & \times P(I(m_1,...,m_n) = i, R(m_1,...,m_n) = j) \end{aligned}$$

Because $(I_1,R_1) \equiv (I(m_1, m_2, m_3,\ldots, m_n), R(m_1, m_2, m_3,\ldots, m_n))$, and $(I_2,R_2) \equiv (I(m_1\text{-}1, m_2\text{+}1, m_3,\ldots, m_n), R(m_1\text{-}1, m_2\text{+}1, m_3,\ldots, m_n))$ have the same joint distribution, it suffices to show that for $m_1 > m_2$ and $k > 0$

$$P(N(m_1,...,m_n) > r | I_1 = 1, R_1 = r - k) \geq P(N(m_1 - 1, m_2 + 1, ..., m_n) > r | I_2 = 1, R_2 = r - k)$$

which is equivalent to showing that

$$P(N(m_1, m_2) > k) \geq P(N(m_1 - 1, m_2 + 1) > k)$$

or, equivalently, that

$$\frac{\binom{m_1}{k} + \binom{m_2}{k}}{\binom{m_1 + m_2}{k}} \geq \frac{\binom{m_1 - 1}{k} + \binom{m_2 + 1}{k}}{\binom{m_1 + m_2}{k}}$$

or, equivalently, that

$$\binom{m}{k}-\binom{m-1}{k}\uparrow m$$

or, equivalently, that

$$\frac{(m-1)!}{(k-1)!(m-k)!}\uparrow m$$

which is immediate. ■

The following proposition gives bounds for *P(N >r)*, as well as a closed form expression when $m_i=m$, $i=1,\ldots,n$.

**Proposition 6**

$$\sum_{j=1}^{n}\frac{\binom{M-m_j}{r}\Big/\binom{M}{r}}{1+\sum_{k\neq j}\binom{M-m_j-m_k}{r}\Big/\binom{M-m_j}{r}}\le P(N>r)\le\sum_{j=1}^{n}\binom{M-m_j}{r}\Big/\binom{M}{r}$$

If all $m_i=m$,

$$P(N>r)=\sum_{i=1}^{n}(-1)^{i+1}\binom{n}{i}\frac{\binom{M-im}{r}}{\binom{M}{r}}$$

**Proof:** Let $A_i$ denote the event that there are no type *i* balls among the first *r* selected. Then

$$P(N>r)=P(A_1\cup A_2\cup....\cup A_n)$$

The right hand inequality follows from Boole's inequality, and the left hand one from the conditional expectation inequality (Ross, 2002). The result when all $m_i=m$, follows from the inclusion-exclusion identity. ■

**Remarks:**

(a) The conditional expectation inequality, stronger than the second moment inequality, states that

$$P(\bigcup_{j=1}^{n}A_j)\ge\sum_{j=1}^{n}\frac{P(A_j)}{1+\sum_{k\neq j}P(A_k|A_j)}$$

(b) Although we could also use the inclusion-exclusion identity to derive an expression for *P(N >r)* in the general case, it would require summing over $2^n$ terms.

(c) When all $m_i = m,$, we can also efficiently compute *P(N >r)* by a recursion. For, $m_i < j$ , let *A*(*i*,*j*) denote a generic random variable having the distribution of the number of additional balls one must remove to have complete set if *j* balls remain in the urn and there are still *i* types that have yet to be collected. Then, with

$$P_k(i, j) = P(A(i, j) \le k)$$

we have the recursion

$$P_k(i, j) = \frac{mi}{j} P_{k-1}(i-1, j-1) + \frac{j-mi}{j} P_{k-1}(i, j-1),\ i > 0$$

with the boundary condition

$$P_k(0, j) = 1; k \ge 0$$

Numerical computation yields the result *P(N>r)=1-P*$_r$*(n,nm)*.

**3. THE DUAL PROBLEM**

The dual problem to the one so far considered is interested in the number of balls that have been removed at the first moment that all the balls of any of the *n* types have been removed.

Now, imagining that we continue removing balls until they are all taken from the urn, the outcome of the experiment for both the original and dual problem is a vector $(i_1,...,i_M)$, where $i_j$ is the type of the $j^{th}$ ball removed. Now, as a function of the outcome, let $N_k(i_1,...,i_M)$ be the number of type *k* balls that have been removed, and let $L(i_1,...,i_M)$ be the last type to have been removed, at the first moment when at least one ball of each type has been removed. Similarly, let $N_k^*(i_1,...,i_M)$ be the number of type *k* balls that have been removed, and let $L^*(i_1,...,i_M)$ be the last type to have been removed, at the first moment that all the balls of any of the *n* types have been removed. It is easy to see that

$$N_k^*(i_1,...,i_M) = \begin{cases} 1 + m_k - N_k(i_M,...,i_1), & \text{if } L(i_1,...,i_M) = k \\ m_k - N_k(i_M,...,i_1), & \text{if } L(i_1,...,i_M) \ne k \end{cases} \quad (1)$$

$$L^*(i_1,...,i_M) = L(i_M,...,i_1) \quad (2)$$

Let $N_k^*$ and $L^*$ be the numbers of type *k* balls removed and the last type removed, respectively, in the dual problem, and let $N_k$ and *L* be the corresponding quantities for the original problem. Because all outcomes are equally likely, it follows from (1) and (2) that

$$P(N_k^* = r, L^* = j) = \begin{cases} P(N_k = 1 + m_k - r, L = j), & \text{if } j = k \\ P(N_k = m_k - r, L = j), & \text{if } j \ne k \end{cases} \quad (3)$$

In addition, it follows from (1) that

$$\sum_{k=1}^{n} N_k^*(i_1,...,i_M) = 1 + M - \sum_{k=1}^{n} N_k(i_M,....,i_1)$$

implying, because all outcomes are equally likely, that

$$P(N^* = j) = P(N = M + 1 - j) \qquad (4)$$

where $N$ and $N^*$ are the numbers of balls removed in the original problem and in the dual problem, respectively.

**Remarks**

(a) The dual problem was previously considered by El-Neweihi et al.(1978), where their main concern was determining the probability mass function of the type of the last ball removed. In addition, they conjectured that $N^*$ was an increasing failure rate random variable. This conjecture was then proven by Ross et al. (1980).

(b) The duality result given by Equation (4) was previously given in Ross et al. (1980).

**REFERENCES**

[1] Adler I., Oren S., Ross S. M. (2003), The coupon collector's problem revisited; *Journal of Applied Probability*, 40; 513-518.

[2] El-Neweihi E., Proschan F., Sethuraman J. (1978), A simple model with applications in structural reliability, extinction of species, inventory depletion and urn sampling; *Advances in Applied Probability*, 10(1); 232-254.

[3] Ross S. M. (2002), Probability models for computer science, Academic Press.

[4] Ross S. M., Shahshahani, M., Weiss G. (1980), On the number of component failures in systems whose component lives are exchangeable, *Mathematics of Operations Research*, 5(3); 358-365.