The Collector's Brotherhood Problem
Using the Newman-Shepp Symbolic Method

Dominique Foata,[1] *Doron Zeilberger*[2]

To Gian-Carlo Rota, in memoriam

Abstract. Further computations are made on the traditional coupon collector's problem when the collector shares his harvest with his younger brothers. When the book of the p-th brother of the collector is completed, the books of the younger brothers have certain numbers of empty spots. On the average, how many? Several answers can be brought to that question.

1. Introduction. This paper on the traditional and recurrent Collector's Problem, that uses the Newman-Shepp method, one of those symbolic approaches dear to Gian-Carlo Rota, is dedicated to his memory.

Suppose that m different coupons are needed for a collector, to complete his picture-book. Each time, that is, at each purchase, he can get a given coupon with probability $1/m$. If he does not have it yet, he sticks it on his book. Otherwise, he gives it to his younger brother. In his turn, the younger brother sticks it on his own book, if the coupon is new for him. Otherwise, the next younger brother will receive it and try to complete his own picture-book, using the same policy. The collector is labelled 0, the younger brother 1, the next younger brother 2, etc.

For $p \ge 0$ let T_p be the time at which the person labelled p completes his book. Also for each $k \ge 1$, $n \ge 1$ and $1 \le i \le m$ let

$$X_{n,i}^{(k)} = \begin{cases} 1, & \text{if coupon } i \text{ has occurred } k \text{ times} \\ & \qquad\qquad \text{up to and including time } n; \\ 0, & \text{otherwise.} \end{cases}$$

Thus, $X_n^{(k)} = \sum\limits_{1\le i\le m} X_{n,i}^{(k)}$ is the number of coupons that have occurred k times (exactly) up to and including time n. For each $n \ge 1$ let $X_n^{(0)}$ be the number of empty spots in the collector's book at time n; by convention $X_0^{(0)} = m$.

AMS 2000 Classification: 05A15, 05E05, 60C05.
Keywords: Coupon collector's problem, Surjection calculus, Newman-Shepp symbolic method, Hyperharmonic numbers.

[1] Supported in part by the NSF, as a consultant on Zeilberger's grant.

[2] Supported in part by the NSF.

For each $p \geq 0$ we have $X_{T_p}^{(p)} = 0$ and for each $k \geq p+1$ the number of coupons that have occurred k times up to and including time T_p is $X_{T_p}^{(k)}$. Consequently, the number of empty spots in the book of the brother labelled $k \geq p+1$ at time T_p is $X_{T_p}^{(p+1)} + X_{T_p}^{(p+2)} + \cdots + X_{T_p}^{(k)}$.

Several methods have been used to derive the probability distribution of T_0 and its expected value $\mathbb{E}[T_0]$. The evaluation $\mathbb{E}[T_0] = m\,H_m$, where H_m is the *harmonic number* $H_m = 1 + \frac{1}{2} + \cdots + \frac{1}{m}$, is a classic (see Feller (1968)). The evaluation $\mathbb{E}[X_{T_0}^{(1)}] = H_m$ is due to Pintacuda (1980) using a method based on the martingale stopping time theorem. Recently, Foata, Han and Lass (2001) have derived the generating function for the vector $(T_0, X_{T_0}^{(1)}, X_{T_0}^{(2)}, \ldots, X_{T_0}^{(k)})$ $(k \geq 1)$ and, when $k \geq 2$, obtained the evaluations $\mathbb{E}[X_{T_0}^{(k)}] = K_m^{(k)}$, where $K_m^{(k)}$ is the *hyperharmonic number* that can be defined by the generating function:

$$(1.1)\qquad \sum_{k\geq 0} K_m^{(k)} t^k = \frac{1}{(1-t/2)(1-t/3)\cdots(1-t/m)}.$$

Now the natural question to ask is the following: what can be said when the stopping time T_0 is replaced by T_p for an arbitrary $p \geq 0$? Back to the early sixties Newman and Shepp (1960) obtained the integral expression

$$(1.2)\qquad \mathbb{E}[T_p] = m \int_0^{+\infty} \bigl(1 - (1 - S_{p+1}(x)e^{-x})^m\bigr)\,dx,$$

where

$$(1.3)\qquad S_{p+1}(x) := 1 + \frac{x}{1!} + \cdots + \frac{x^p}{p!} \quad (p \geq 0).$$

The purpose of this paper is to make use of the *symbolic method* developed by Newman and Shepp (*op. cit.*) to obtain the following formulas, all of them expressed in terms of *integrals*. For each $0 \leq p < r$ consider the generating functions:

$$(1.4)\qquad H_{T_p}(t) := \sum_{i\geq 0} \mathrm{P}\{T_p > i\}\, t^i;$$

$$(1.5)\qquad G_{T_p}(t) := \sum_{i\geq 0} \mathrm{P}\{T_p = i\}\, t^i;$$

$$(1.6)\quad G_{T_p,X}(t, \mathbf{s}) := \sum_{n,n_1,\ldots,n_r} \mathrm{P}\{T_p = n, X_{T_p}^{(p+1)} = n_{p+1}, \ldots, X_{T_p}^{(r)} = n_r\} \times t^n s_{p+1}^{n_{p+1}} \cdots s_r^{n_r};$$

$$(1.7)\qquad G_X^{(p)}(t) := \sum_{k\geq p+1} \mathbb{E}[X_{T_p}^{(k)}]\, t^k.$$

We then establish the formulas:

$$(1.8)\quad H_{T_p}(t) = m\int_0^{+\infty}\bigl(e^{-mx(1-t)} - (e^{-x(1-t)} - S_{p+1}(tx)e^{-x})^m\bigr)\,dx;$$

$$(1.9)\quad G_{T_p}(t) = m(1-t)\int_0^{+\infty}\bigl(e^{-x(1-t)} - e^{-x}S_{p+1}(tx)\bigr)^m\,dx;$$

$$(1.10)\quad G_{T_p,X}(t,\mathbf{s}) = \frac{m s_{p+1} t^{p+1}}{p!}\int_0^{+\infty} x^p e^{-mx} \times \Bigl(e^{xt} - S_{r+1}(xt) + \frac{(xt)^{p+1}}{(p+1)!}s_{p+1} + \cdots + \frac{(xt)^r}{r!}s_r\Bigr)^{m-1}\,dx;$$

$$(1.11)\quad G_X^{(p)}(t) = t^{p+1} + m(m-1)\int_0^{+\infty} e^{-mx}\frac{x^p}{p!}\bigl(e^{xt} - S_{p+1}(xt)\bigr)\bigl(e^x - S_{p+1}(x)\bigr)^{m-2}\,dx.$$

By expanding the different integrands using the multinomial theorem we can express those integrals as finite sums over $\mathbb{N}^{r+2}$ for some r. For instance, (1.10) may be rewritten as

$$(1.12)\quad G_{T_p,X}(t,\mathbf{s}) = \frac{s_{p+1}t}{p!}\sum\binom{m-1}{a, b_0, \ldots, b_r}(-1)^{b_0+\cdots+b_r}\Biggl(\prod_{i=0}^{p}\Bigl(\frac{1}{i!}\Bigr)^{b_i}\Biggr) \times \Biggl(\prod_{i=p+1}^{r}\Bigl(\frac{s_i-1}{i!}\Bigr)^{b_i}\Biggr)\frac{(t/m)^{p+1\cdot b_1+\cdots+r\cdot b_r}}{(1-at/m)^{1+p+1\cdot b_1+\cdots+r\cdot b_r}}(p+\sum_{i=1}^{r} ib_i)!,$$

that specializes to the formula derived by Foata et al. (2001) for the case $p=0$. However the main difference between the latter case and the general case $p \geq 1$ is the fact that when $p=0$ identity (1.12) can directly lead to the *explicit* evaluations $\mathbb{E}[X_{T_0}^{(1)}] = H_m$ and $\mathbb{E}[X_{T_0}^{(k)}] = K_m^{(k)}$ $(k \geq 2)$ (the hyperharmonic number), while for $p \geq 1$ the *integral form* that can be obtained from (1.10) has no further simplification.

In section 2 we discuss certain salient features of Surjection Calculus. In section 3 we recall the *symbolic method* developed by Newman and Shepp (*op. cit.*) and use it to prove identities (1.8) and (1.9). Identities (1.10) and (1.11) are proved in sections 4 and 5, respectively. We conclude the paper by some remarks on the holonomic aspects of the above integral identities.

Notice that the asymptotics of $\mathbb{E}[T_p]$ for an arbitraary $p \geq 0$ have been derived by Newman and Shepp (*op. cit.*) and the asymptotic probability distribution of T_p has been calculated by Erdős and Rényi (1961). More recent work along those lines is due to Wilf (2001).

2. Surjection calculus. Let $(t, s_1, s_2, \dots)$ be a sequence of commuting variables. If f maps a finite set A of integers into a (finite) set B of integers, let $\nu_i(f)$ denote the number of elements $j \in B$ such that the preimage $f^{-1}(j)$ is of cardinality i. Then, the *weight* $\pi(f)$ of f is defined by: $\pi(f) := \prod\limits_{i\ge 1} s_i^{\nu_i(f)}$; so that the relations $\sum\limits_i \nu_i(f) = \#B$ et $\sum\limits_i i\,\nu_i(f) = \#A$. hold.

For each integer n let $[n] := \{1, 2, \dots, n\}$ and for each pair of finite sets (A, B) let $\mathrm{Surj}_{\ge p+1}(A, B)$ denote the set of all *surjections* $f : A \to B$ such that for all $j \in B$ the inequality $\#f^{-1}(j) \ge p + 1$ holds. The following identity is the basic identity in Surjection Calculus (see, e.g., Foata (1974), p. 60):

$$\Bigl(\sum_{i\ge p+1} s_i \frac{t^i}{i!}\Bigr)^m = \sum_{n\ge (p+1)m} \frac{t^n}{n!} \sum_{f\in \mathrm{Surj}_{\ge p+1}([n],[m])} \pi(f). \tag{2.1}$$

This implies, when all the s_i's are equal to 1,

$$(e^t - S_{p+1}(t))^m = \sum_{n\ge (p+1)m} \frac{t^n}{n!} \#\,\mathrm{Surj}_{\ge p+1}([n],[m]). \tag{2.2}$$

Next consider the subset $A(n, m)$ of all surjections f in $\mathrm{Surj}_{\ge p+1}([n], [m])$ such that if $f(n) = j$ then $\#f^{-1}(j) = p + 1$. In other words, all values are taken at least $p+1$ times, but the value taken at n occurs exactly p times in $(f(1), f(2), \dots, f(n-1))$. Then, identity (2.1) implies

$$\sum_{n\ge m(p+1)} \frac{t^{n-1}}{(n-1)!} \sum_{f\in A(n,m)} \pi(f) = m s_{p+1} \frac{t^p}{p!} \Bigl(\sum_{i\ge p+1} s_i \frac{t^i}{i!}\Bigr)^{m-1}. \tag{2.3}$$

Let $0 \le p < r$, $\mathbf{n} := (n_{p+1}, n_{p+2}, \dots, n_r)$ and let $A(n, m; \mathbf{n})$ be the subset of all surjections f in $A(n, m)$ such that

$$\nu_i(f) = n_i \quad \text{for } i = p + 1, p + 2, \dots, r.$$

When $s_i := 1$ for all $i \ge r + 1$, identity (2.3) rewrites as:

$$\begin{aligned}\sum_{n\ge m(p+1)} \frac{t^{n-1}}{(n-1)!} \sum_{\mathbf{n}} \#A(n, m; \mathbf{n}) \prod_{p+1\le i\le r} s_i^{n_i}\qquad\qquad&\\ = m s_{p+1} \frac{t^p}{p!} \Bigl(e^t - S_{r+1}(t) + s_{p+1} \frac{t^{p+1}}{(p+1)!} + \dots + s_r \frac{t^r}{r!}\Bigr)^{m-1}&.\end{aligned} \tag{2.4}$$

Now let i, j be two distinct integers in $[m]$ and for each $n \ge (p+1)m$ let $B(n-1, i, j)$ be the set of surjections $f : [n-1] \to [m]$ with the following properties: with $I(f) := f^{-1}(i)$, $J(f) := f^{-1}(j)$ then $\#J(f) = p$ and the restriction of f to $[n-1]\setminus J(f)$ belongs to $\mathrm{Surj}_{\geq p+1}([n-1]\setminus J(f), [m]\setminus\{j\})$. Then form the generating polynomial

$$F_{n-1}(t) = \sum_{f\in B(n-1,i,j)} t^{\#I(f)}. \tag{2.5}$$

Still using (2.1) a little reflection shows that the following identity holds:

$$\sum_{n\geq(p+1)m} \frac{u^{n-1}}{(n-1)!} F_{n-1}(t) = \frac{u^p}{p!}\big(e^{ut} - S_{p+1}(ut)\big)\big(e^u - S_{p+1}(u)\big)^{m-2}. \tag{2.6}$$

3. The Newman-Shepp symbolic method. Let $m \geq 1$ and consider the algebra of power series in the variables u_1, u_2, $\ldots$, u_n. If $a = a(u_1, \ldots, u_n)$ is such a series, write it as

$$a(u_1, \ldots, u_n) = \sum_{i\geq 0} \frac{1}{i!}\, a_i(u_1, \ldots, u_n),$$

where $a_i(u_1, \ldots, u_n)$ is the sum of all monomials of degree i in a. Now define

$$N_m\, a(u_1, \ldots, u_n) := \sum_{i\geq 0} \frac{1}{m^i}\, a_i(u_1, \ldots, u_n),$$

so that the *exponential* normalization has be replaced by the *power* normalization. We next reproduce the integral expression for $N_m\, a(u_1, \ldots, u_n)$ as derived by Newman and Shepp (*op. cit.*).

Lemma 3.1. *We have the identity:*

$$N_m\, a(u_1, \ldots, u_n) = m \int_0^{+\infty} e^{-mx}\, a(xu_1, \ldots, xu_n)\, dx.$$

For each $n = 1, 2, \ldots$ let Y_n be the label of the coupon which is obtained at time n. By assumption, the random variables Y_n $(n = 1, 2, \ldots)$ are assumed to be independent and uniformly distributed on the coupon label set $\{1, 2, \ldots, m\}$. The event $\{T_p > n\}$ is realized if and only if the mapping $f : i \mapsto Y_i$ $(1 \leq i \leq n)$ belongs to the set $\mathrm{Surj}^c_{\geq p+1}([n], [m]) := [m]^n \setminus \mathrm{Surj}_{\geq p+1}([n], [m])$. Hence $\mathrm{P}\{T_p > n\} = \#\,\mathrm{Surj}^c_{\geq p+1}([n], [m])/m^n$. But it follows from (2.2) that

$$e^{mt} - (e^t - S_{p+1}(t))^m = \sum_{n\geq 0} \frac{t^n}{n!} \#\,\mathrm{Surj}^c_{\geq p+1}([n], [m]),$$

so that

$$N_m\left(e^{mt}-(e^t-S_{p+1}(t))^m\right)=\sum_{n\ge 0}\frac{t^n}{m^n}\#\operatorname{Surj}^c_{\ge p+1}([n],[m])$$
$$=\sum_{n\ge 0}\mathrm{P}\{T_p>n\}t^n=H_{T_p}(t).$$

Hence, by Lemma 3.1

$$H_{T_p}(t)=m\int_0^{+\infty}e^{-mx}(e^{mtx}-\bigl(e^{tx}-S_{p+1}(tx))^m\bigr)\,dx,$$

which reduces to (1.8). To get (1.9) we make use of the traditional relation $G_{T_p}(t)=1-(1-t)\,H_{T_p}(t)$. Finally, when we plug $t=1$ in (1.8), we recover formula (1.2) already derived by Newman and Shepp (*op. cit.*).

4. The multivariable generating function. Keep the same notations for the sequence (Y_n) $(n=1,2,\ldots)$ of coupon labels that occur at time $n=1,2,\ldots$ and for the variables $X_n^{(k)}$ defined in the introduction. Then, the event $\{T_p=n, X_{T_p}^{(p+1)}=n_{p+1},\ldots,X_{T_p}^{(r)}=n_r\}$ is realized if and only if the mapping $f:i\mapsto Y_i$ $(1\le i\le n)$ belongs to the set $A(n,m;\mathbf{n})$ introduced in (2.4). Hence, $\mathrm{P}\{T_p=n, X_{T_p}^{(p+1)}=n_{p+1},\ldots,X_{T_p}^{(r)}=n_r\}=\#A(n,m;\mathbf{n})/m^n$. It follows from (2.4) that

$$N_m\Bigl(ms_{p+1}\frac{t^p}{p!}\Bigl(e^t-S_{r+1}(t)+s_{p+1}\frac{t^{p+1}}{(p+1)!}+\cdots+s_r\frac{t^r}{r!}\Bigr)^{m-1}\Bigr)$$
$$=\sum_{n\ge m(p+1)}\frac{t^{n-1}}{m^{n-1}}\sum_{\mathbf{n}}\#A(n,m;\mathbf{n})\prod_{p+1\le i\le r}s_i^{n_i},$$

so that, by multiplying both members by t/m,

$$t\,N_m\Bigl(s_{p+1}\frac{t^p}{p!}\Bigl(e^t-S_{r+1}(t)+s_{p+1}\frac{t^{p+1}}{(p+1)!}+\cdots+s_r\frac{t^r}{r!}\Bigr)^{m-1}\Bigr)$$
$$=\sum_{n\ge m(p+1)}\frac{t^n}{m^n}\sum_{\mathbf{n}}\#A(n,m;\mathbf{n})\prod_{p+1\le i\le r}s_i^{n_i}$$
$$=\sum_{n,\mathbf{n}}\mathrm{P}\{T_p=n,X_{T_p}^{(p+1)}=n_{p+1},\ldots,X_{T_p}^{(r)}=n_r\}t^n s_{p+1}^{n_{p+1}}\cdots s_r^{n_r}$$
$$=G_{T_p,X}(t,\mathbf{s}).$$

This establishes identity (1.10) by using again Lemma 3.1.

5. A generating function for the expected times. We may write

$$\begin{aligned}
G_X^{(p)}(t) &:= \sum_{k\ge p+1} \mathbb{E}[X_{T_p}^{(k)}]\, t^k = \sum_{k\ge p+1}\sum_{1\le i\le m} \mathbb{E}[X_{T_p,i}^{(k)}]\, t^k \\
&= \sum_{k\ge p+1}\sum_{1\le i\le m} \mathrm{P}\{X_{T_p,i}^{(k)}=1\}\, t^k \\
&= \sum_{k\ge p+1}\sum_{1\le i\le m}\sum_{n\ge (p+1)m}\sum_{1\le j\le m} \mathrm{P}\{X_{n,i}^{(k)}=1, T_p=n, Y_n=j\}\, t^k \\
&= t^{p+1} + \sum_{n\ge (p+1)m}\sum_{i}\sum_{j\ne i}\sum_{k\ge p+1} \mathrm{P}\{X_{n-1,i}^{(k)}=1, T_p=n, Y_n=j\}\, t^k.
\end{aligned}$$

Now the event $\{X_{n-1,i}^{(k)}=1, T_p=n, Y_n=j\}$ is realized if and only if the mapping $f: k\mapsto Y_k$ $(1\le k\le n-1)$ belongs to the set $B(n-1,i,j)$ introduced in (2.5)–(2.6) and if $Y_n=j$. Hence, keeping the same notations,

$$\sum_{k\ge p+1} \mathrm{P}\{X_{n-1,i}^{(k)}=1, T_p=n, Y_n=j\}t^k = \frac{1}{m^n}F_{n-1}(t),$$

so that

$$G_X^{(p)}(t) = t^{p+1} + m(m-1)\sum_{n\ge (p+1)m} \frac{1}{m^n}F_{n-1}(t).$$

With the introduction of a new variable u define:

$$G_X^{(p)}(t,u) := t^{p+1} + (m-1)\sum_{n\ge (p+1)m} \frac{u^{n-1}}{m^{n-1}}F_{n-1}(t).$$

Then

$$\begin{aligned}
G_X^{(p)}(t,u) &= N_m\Bigl(t^{p+1} + (m-1)\sum_{n\ge (p+1)m} \frac{u^{n-1}}{(n-1)!}F_{n-1}(t)\Bigr) \\
&= N_m\Bigl(t^{p+1} + (m-1)\frac{u^p}{p!}\bigl(e^{ut}-S_{p+1}(ut)\bigr)\bigl(e^{u}-S_{p+1}(u)\bigr)^{m-2}\Bigr)
\end{aligned}$$

by using (2.6). Next applying Lemma 3.1 we get:

$$\begin{aligned}
&G_X^{(p)}(t,u) = t^{p+1} \\
&+ m(m-1)\int_0^{+\infty} e^{-mx}\frac{(ux)^p}{p!}\bigl(e^{uxt}-S_{p+1}(uxt)\bigr)\bigl(e^{ux}-S_{p+1}(ux)\bigr)^{m-2}\,dx,
\end{aligned}$$

which reduces to (1.11) by making $u=1$.

6. Concluding Remarks. Identity (1.1) together with the evaluations $\mathbb{E}[X_{T_0}^{(1)}] = H_m$ and $\mathbb{E}[X_{T_0}^{(k)}] = K_m^{(k)}$ $(k \geq 2)$ can be rewritten as

$$(6.1)\quad G_X^{(0)}(t) = \sum_{k\geq 1} \mathbb{E}[X_{T_0}^{(k)}]t^k = -1+t+\frac{1}{(1-t/2)(1-t/3)\cdots(1-t/m)}.$$

That identity, derived by Foata et al. (*op. cit.*), can also be obtained from its integral form (1.11). When $p = 0$, the factor x^p in the integrand vanishes. With the change of variables $u = e^{-x}$ the integral in (1.11) reduces to the evaluation of two bêta integrals:

$$\begin{aligned} G_X^{(0)}(t) &= t + m(m-1)\int_0^1 (u^{-t+1} - u)(1-u)^{m-2}\,du \\ &= t + m(m-1)\Bigl(\frac{\Gamma(2-t)\Gamma(m-1)}{\Gamma(m-t+1)} - \frac{\Gamma(2)\Gamma(m-1)}{\Gamma(m+1)}\Bigr) \\ &= t + \frac{m!}{(m-t)(m-t-1)\cdots(2-t)} - 1 \\ &= -1 + t + \frac{1}{(1-t/2)(1-t/3)\cdots(1-t/m)}. \quad \square \end{aligned}$$

Unfortunately, for $p > 0$, it does not seem to be possible to evaluate the integral in closed form, since then the integral representation is *hybrid*, involving a polynomial in the variable and its exponential. Hence it lies outside the jurisdiction of the *holonomic paradigm*, and in general there is no reason why the integral should evaluate to something 'nice', and indeed it, most probably, does not!

Acknowledgements. We owe to the search sagacity of Herb Wilf to have found track of the paper by Newman and Shepp. Those authors had entitled their paper "Double Dixie Cup Problem" without reference to the traditional Coupon Collector lore, which explains our difficulty to locate it through *MathSciNet*. We should also like to thank the anonymous referee for his careful reading of the paper and his useful remarks.

REFERENCES

Erdős, Paul; Rényi, Alfred (1961). On a classical problem of probability theory, *Magyar. Tud. Akad. Mat. Kutató Int. Kőzl.*, **6**, p. 215–220.

Feller, William (1968). *An Introduction to Probability Theory and its Applications*, Third edition, vol. 1. John Wiley & Sons, New York.

Foata, Dominique (1974). *La série génératrice exponentielle dans les problèmes d'énumération.* Les Presses de l'Université de Montréal, Montréal.

Foata, Dominique; Han Guo-Niu; Lass, Bodo (2001). Les nombres hyperharmoniques et la fratrie du collectionneur de vignettes, *Sém. Lothar. Combin.*, **47**, B47a, 2001, 20 pages [`http://mat.univie.ac.at/~slc/`].

Newman, Donald J.; Shepp, Lawrence (1960). The Double Dixie Cup Problem, *Amer. Math. Monthly*, **67**, p. 58–61.

Pintacuda N. (1980). Coupon Collectors via the Martingales, *Boll. Un. Mat. Ital. A*, **17**, p. 174–177.

Wilf, Herb (2001). Some finite structures of the coupon collector's problem. Preprint.

Dominique Foata
Département de mathématique
Université Louis Pasteur
7, rue René-Descartes
F-67084 Strasbourg, France
`foata@math.u-strasbg.fr`

Doron Zeilberger
Department of Mathematics
Hill Center-Busch Campus
Rutgers, The State University of New Jersey
110 Frelinghuysen Rd
Piscataway, NJ 08854-8019, USA
`zeilberg@math.rutgers.edu`