Coupon Collecting with Quotas

Russell May
150 University Blvd., UPO 701
Morehead State University
Morehead, KY 40351-1689, USA
r.may@moreheadstate.edu

Submitted: April 7, 2008; Accepted: Jul 28, 2008; Published: Aug 18, 2008
Mathematics Subject Classifications: 05A15, 60C05

Abstract

We analyze a variant of the coupon collector's problem, in which the probabilities of obtaining coupons and the numbers of coupons in a collection may be non-uniform. We obtain a finite expression for the generating function of the probabilities to complete a collection and show how this generalizes several previous results about the coupon collector's problem. Also, we provide applications about computational complexity and approximation.

1 Introduction

Soft drink manufacturers have popularized the "under-the-cap game," in which they imprint a letter of a payoff word (usually the name of the manufacturer itself) underneath bottle caps and dispense bottles of the soft drink randomly. Consumers then buy bottle after bottle of the soft drink, hoping to collect enough letters to spell out the payoff word. Discerning consumers might wonder how many bottles they would expect to purchase in order to spell out the payoff word and win the game. If the letters in the payoff word are distinct, like in *Sprite®*, and the letters are distributed uniformly, this problem is the same as the classic coupon collector's problem, in which coupons of d kinds are randomly dispensed, and collectors ask how many coupons they must obtain on average to form a complete set of at least one of each kind. A classic argument shows that on average a collector must obtain $d(1 + \frac{1}{2} + \cdots + \frac{1}{d})$ coupons to form a complete set.

Generalizations of the coupon collector's problem date back to at least 1934, when von Schelling in [6] (and re-published in [7]) computed the expected number of coupons to obtain a complete collection under the condition that the probabilities of obtaining a coupon could be non-uniform. Then in 1960, Newman and Shepp in their well-known "The Double Dixie Cup Problem" [4] generalized the coupon collector's problem to find the expected number of coupons to obtain an arbitrary number of complete sets, but with a uniform distribution of the coupons. Wilf and Myers in [3] re-derived the result of Newman and Shepp, but with a generating function of just one variable instead of several. Further generalizations of the coupon collector's problem are still a fruitful source of contemporary research (see for instance [1] or [2]). Surely, one reason for the problem's continued popularity is the uncanny way in which infinite series related to the problem turn out to be expressible in finite terms. This note continues on that theme.

We consider the under-the-cap game with a payoff word having repeated letters, for example, as in *Dr. Pepper*®. Each of the letters D, E, P, and R must be collected a certain number of times, called its *quota*, which in general may be greater than one and may vary from letter to letter. Also, the probabilities of obtaining the letters may be non-uniform. For example, we could model an "under-the-cap" game for Dr. Pepper as follows:

Letter	Quota	Probability
D	1	.25
E	2	.25
P	3	.15
R	2	.35

The general problem that we solve, the "coupon collector's problem with quotas," is for a payoff word with letters in the set L which appear with probabilities $\vec{p} = \langle p_\ell : \ell \in L\rangle$ and quotas $\vec{q} = \langle q_\ell : \ell \in L\rangle$ to find the expected number $\langle T_{\vec{p},\vec{q}}\rangle$ of bottles a consumer must purchase in order to spell out the payoff word, i.e., to obtain at least q_ℓ copies of ℓ for each letter ℓ in L. The only assumptions about the succession of letters are that the letters on the bottles are independent and that probabilities of letters under each bottle are identically distributed.

To fix notation, for non-negative integers n and r let $\binom{n}{r}$ denote the binomial coefficient $\frac{n(n-1)\cdots(n-r+1)}{r!}$. Likewise, if D is a linear operator, let $\binom{D}{r}$ denote the operator $\frac{D(D-1)\cdots(D-r+1)}{r!}$. If $\vec{r}$ is a k-tuple of non-negative integers with sum n, let $\binom{n}{\vec{r}}$ denote the multinomial coefficient $\frac{n!}{r_1!r_2!\cdots r_k!}$. Lastly, let $T_n(x)$ be $1+x+\frac{x^2}{2!}+\cdots+\frac{x^{n-1}}{(n-1)!}$, the n^{th} order Taylor polynomial of the exponential function.

2 A Generating Function for Winning the Game

In this section we find an expression with finitely many terms for the generating function of the sequence of probabilities a_n that a collection of letters is completed on the n^{th} bottle. This calculation closely follows the style of section 3 of [3], but generalizes the main result there (Theorem 2, equation 35) to non-uniform probabilities and quotas. For an excellent primer on generating functions, whose basic results are used here, see [9].

To win the under-the-cap game on the n^{th} bottle for a collection whose letters L have probabilities $\langle p_\ell : \ell \in L\rangle$ and quotas $\langle q_\ell : \ell \in L\rangle$, one letter ℓ must meet its quota q_ℓ on the n^{th} bottle, meaning that exactly $q_\ell - 1$ appearances of ℓ must have occurred somewhere among the first $n-1$ bottles, and the rest of the letters must have met or exceeded their quotas on the other $n - q_\ell$ bottles. Evidently,

$$a_n = \sum_{\ell \in L} p_\ell \binom{n-1}{q_\ell - 1} {p_\ell}^{q_\ell - 1} \sum_{\vec{r} \in Q^{n-q_\ell}_{L-\{\ell\}}} \binom{n - q_\ell}{\vec{r}} \prod_{k \in L - \{\ell\}} {p_k}^{r_k},$$

where Q^i_M consists of the finite sequences $\vec{r}$ of integers indexed by the letters in M such that the sum of the integers in $\vec{r}$ is i and $r_m \geq q_m$ for each m in M. We define the ordinary generating function of this sequence of probabilities, $P_{\vec{p},\vec{q}}(x) = \sum_{n \geq 0} a_n x^n$. The goal of this section is to find a finite sum for this generating function.

As an intermediate step, consider the ordinary generating function

$$O_\ell(x) = \sum_{n \geq 0} x^n \sum_{\vec{r} \in Q^n_{L-\{\ell\}}} \binom{n}{\vec{r}} \prod_{k \in L-\{\ell\}} {p_k}^{r_k}$$

and the corresponding exponential generating function

$$E_\ell(x) = \sum_{n \geq 0} \frac{x^n}{n!} \sum_{\vec{r} \in Q^n_{L-\{\ell\}}} \binom{n}{\vec{r}} \prod_{k \in L-\{\ell\}} {p_k}^{r_k}.$$

In terms of the O_ℓ's we can rewrite the original generating function as

$$P_{\vec{p},\vec{q}}(x) = \sum_{\ell \in L} {p_\ell}^{q_\ell} \binom{x\frac{\partial}{\partial x} - 1}{q_\ell - 1} \left[x^{q_\ell} O_\ell(x)\right]. \tag{1}$$

By the exponential formula, we can write each E_ℓ as a finite product, namely

$$E_\ell(x) = \prod_{k \in L-\{\ell\}} \left(e^{p_k x} - T_{q_k}(p_k x)\right). \tag{2}$$

As usual, O_ℓ can be obtained from E_ℓ by taking a Laplace transform, specifically

$$O_\ell(x) = \frac{1}{x} \int_0^\infty e^{-t/x} E_\ell(t)\, dt. \tag{3}$$

Substituting equation 2 into equation 3 and then into equation 1, we have

$$P_{\vec{p},\vec{q}}(x) = \sum_{\ell \in L} {p_\ell}^{q_\ell} \int_0^\infty \binom{x\frac{\partial}{\partial x} - 1}{q_\ell - 1} x^{q_\ell - 1} e^{-t/x} E_\ell(t)\, dt.$$

Noting that $\binom{x\frac{\partial}{\partial x}-1}{q_\ell - 1}\left[x^{q_\ell-1}e^{-t/x}\right] = \frac{t^{q_\ell - 1}}{(q_\ell - 1)!} e^{-t/x}$, we get a convenient form for the generating function

$$P_{\vec{p},\vec{q}}(x) = \sum_{\ell \in L} \frac{{p_\ell}^{q_\ell}}{(q_\ell - 1)!} \int_0^\infty t^{q_\ell - 1} e^{-t/x} \prod_{k \in L-\{\ell\}} \left(e^{p_k t} - T_{q_k}(p_k t)\right) dt, \tag{4}$$

which is a sum with at most $|L| \cdot \prod_{\ell \in L}(q_\ell + 1)$ terms, as desired.

3 Reduction to Previous Results

Equation 4 generalizes Theorem 2 of section 3 in [3], which describes the generating function of the coupon collector's problem for n copies of d coupons distributed uniformly to non-uniform probabilities and quotas. One immediate consequence of equation 4 is an expression with finitely many terms for the expected number of bottles $\langle T_{\vec{p},\vec{q}}\rangle$ needed to win the under-the-cap game,

$$\begin{aligned}\langle T_{\vec{p},\vec{q}}\rangle &= P'_{\vec{p},\vec{q}}(1)\\ &= \sum_{\ell\in L}\frac{{p_\ell}^{q_\ell}}{(q_\ell-1)!}\int_0^\infty t^{q_\ell}e^{-t}\prod_{k\in L-\{\ell\}}\left(e^{p_k t}-T_{q_k}(p_k t)\right)dt. \end{aligned} \tag{5}$$

By expanding the product of sums in equation 5 and noting $\int_0^\infty t^q e^{-pt}\,dt = q!/p^{q+1}$, we have

$$\langle T_{\vec{p},\vec{q}}\rangle = \sum_{M\in\mathcal{P}'(L)}(-1)^{|M|+1}\sum_{\vec{r}\in Q_M^<}\frac{\binom{\sum_{\ell\in M} r_\ell}{\vec{r}}\left(\prod_{\ell\in M} p_\ell^{r_\ell}\right)}{\left(\sum_{\ell\in M} p_\ell\right)^{1+\sum\limits_{\ell\in M} r_\ell}}, \tag{6}$$

where $\mathcal{P}'(L)$ denotes the collection of non-empty subsets of letters in L and $Q_M^<$ denotes the collection of finite sequences $\vec{r}$ of integers indexed by the letters in M such that $0 \le r_\ell < q_\ell$ for each $\ell \in M$. This generalizes von Schelling's result in [6] about the expected number of coupons to obtain a collection of at least one coupon in the non-uniform probability case to non-uniform quotas. A numerical calculation based on equation 6 yielded that the expected number of bottles to win the Dr. Pepper under-the-cap game described in the introduction is approximately 21.156 bottles and to win twice is 40.625 bottles.

For the remainder, we concentrate on the special case of uniform probabilities and quotas. In other words, we suppose the payoff word consists of d distinct letters distributed uniformly and that a collector must obtain n copies of each letter. We let $\langle T_{d,n}\rangle$ be the expected number of bottles necessary to obtain this collection. Then, equation 5 reduces to

$$\langle T_{d,n}\rangle = \frac{d}{(n-1)!}\int_0^\infty e^{-x}x^n\left(1-e^{-x}T_n(x)\right)^{d-1}dx. \tag{7}$$

For the case of only two letters ($d=2$) equation 7 further reduces to

$$\langle T_{2,n}\rangle = 2n\left(1+\binom{2n}{n}4^{-n}\right),$$

which is equivalent to a result of Nishi and Nomakuchi in [5].

4 Computational Complexity

Numerical computation of $\langle T_{d,n}\rangle$ based on equation 7 is computationally infeasible since direct expansion of the integrand in this equation leads to $O(n^d)$ terms. However, there is a more efficient algorithm to compute $\langle T_{d,n}\rangle$, which we now describe. First, by applying integration by parts $d-1$ times to the integral in equation 7, we get

$$\begin{aligned}\langle T_{d,n+1}\rangle &= \frac{d(n+1)}{(d!)^n}\sum_{m_2=0}^{n+1}\binom{n+m_2}{n}\left(\frac{1}{2}\right)^{m_2}\cdots \\ &\qquad\sum_{m_r=0}^{n+m_{r-1}}\binom{n+m_r}{n}\left(\frac{r-1}{r}\right)^{m_r}\cdots \\ &\qquad\qquad\sum_{m_d=0}^{n+m_{d-1}}\binom{n+m_d}{n}\left(\frac{d-1}{d}\right)^{m_d}. \end{aligned} \tag{8}$$

The form of the nested sum in equation 8 is special because the terms in the r^{th} sum only depend on m_r, not the previous $m_2,\ldots,m_{r-1}$. Therefore, the entire sum can be computed in $\sum_{r=2}^{d}\max(m_r)\le\sum_{r=2}^{d} nr = O(nd^2)$ steps. For example, using equation 8 a numerical computation showed that to the nearest integer $\langle T_{100,100}\rangle$ is 12690, whereas a computation of $\langle T_{100,100}\rangle$ from equation 7 with 10^{200} terms would be infeasible.

5 Asymptotic Approximation of Expectation

Asymptotic approximation of the expectation $\langle T_{d,n}\rangle$ for large d or large n is useful for both computational and theoretical reasons. We derive an asymptotic approximation of $\langle T_{d,n}\rangle$ beginning with its representation in equation 7 and making a sequence of four estimates:

$$\langle T_{d,n+1}\rangle = d^2\int_0^\infty x e^{-zd}\,dz \tag{9}$$

$$\approx d^2\int_0^\infty \big(n+\sqrt{2n}\,\operatorname{erfc}^{-1}(z)\big)e^{-zd}\,dz \tag{10}$$

$$\approx d^2\int_0^\infty \big(n+\big(-2n\log(z\sqrt{2\pi})\big)^{\frac{1}{2}}\big)e^{-zd}\,dz \tag{11}$$

$$\approx d\big(n+\big(2n\log(d/\sqrt{2\pi})\big)^{\frac{1}{2}}\big). \tag{12}$$

In equation 9 we make the substitution $e^{-z} = 1-e^{-x}T_{n+1}(x)$ into the integral from equation 7. Equation 10 follows from an application of Laplace's method to approximate $\int_0^x e^{-t}t^n\,dt = n!(1-T_{n+1}(x))$. As a consequence, we have a result first due to Szegö in [8] that an asymptotic approximation for large n of a solution for x in this substitution is $x\approx n+\sqrt{2n}\,\operatorname{erfc}^{-1}(z)$, where erfc denotes the complementary error function $x\mapsto \frac{2}{\sqrt{\pi}}\int_x^\infty e^{-t^2}dt$. To get equation 11, we use the first-order approximation $\operatorname{erfc}(x)\approx\frac{e^{-x^2}}{\sqrt{\pi}\,x}$ so that $\operatorname{erfc}^{-1}(z)\approx\big(-\log(z\sqrt{2\pi})\big)^{\frac{1}{2}}$. In equation 12, we use an approximation of the Laplace transform of a power of a logarithm, $\int_0^c(-\log x)^\mu e^{-kx}\,dx\approx\frac{(\log k)^\mu}{k}$ for large k, (see, for instance, theorem II.2.2 of [10]). As a comparison of results, the asymptotic approximation in equation 12 adds a term proportional to $\sqrt{n}$ that is not included in a

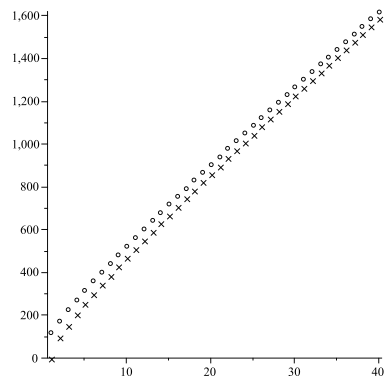

Figure 1: Comparison of approximate ($\times$) and exact ($\circ$) values of $\langle T_{d,n}\rangle$ for $d = 30$ letters and $n = 1$ to $n = 40$ copies.

similar calculation in [3] (equation 43). For a numerical example, our approximation gives $\langle T_{100,100}\rangle \approx 12601$, less than one percent off the exact figure computed from equation 8. Also, figure 1 shows nice agreement between exact and approximate results.

Due to the factor of $\sqrt{n}$ in equation 12, for each d the graph of $n \mapsto \langle T_{d,n}\rangle$ is convex down, as intuition about the expected number of bottles would suggest. More generally, the forward differences of $\langle T_{d,n}\rangle$ with d fixed, defined by $\triangle^0\langle T_{d,n}\rangle = \langle T_{d,n}\rangle$ and $\triangle^{r+1}\langle T_{d,n}\rangle = \triangle^r\langle T_{d,n+1}\rangle - \triangle^r\langle T_{d,n}\rangle$, depend only on the parity of r, namely $\mathrm{sign}(\triangle^r\langle T_{d,n}\rangle) = \mathrm{sign}(\triangle^r\sqrt{n}) = (-1)^{r+1}$ for $r \geq 1$. Oddly enough, this property does not always hold in the non-uniform case. Consider the number of bottles $\langle T_{\vec{p},n\,\vec{q}_0}\rangle$ needed to collect the payoff word n times, i.e., q_ℓ is n times the number of occurrences of letter ℓ in the payoff word. In the Dr. Pepper under-the-cap game described in the introduction, a numerical calculation showed that $n \mapsto \langle T_{\vec{p},n\,\vec{q}_0}\rangle$ is not even convex, contrary to the pattern in the uniform case even for $r = 2$.

References

[1] Adler, I., Oren, S., and Ross, S. (2003). The Coupon Collector's Problem Revisited, *J. Appl. Probab.*, **40**, no. 2, 513-518.

[2] Foata, D. and Zeilberger, D. (2002). The Collectors Brotherhood Problem Using the Newman-Shepp Symbolic Method, *Algebra Universalis*, **49**, 387–395.

[3] Myers, A. and Wilf, H. (2003). Some New Aspects of the Coupon-Collectors Problem, *SIAM Journal on Discrete Mathematics*, **17**, no. 1, 1–17.

[4] Newman, D. and Shepp, L. (1960). The Double Dixie Cup Problem, Amer. Math. Monthly, **67**, 58-61.

[5] Nishi, A. and Nomakuchi, K. (1986). A Note on the Coupon Collector's Problems, *Journal of the Faculty of Education, Saga University*, **33**, no. 2, 185–190.

[6] von Schelling, H., (1934). Auf der Spur des Zufalls. *Deutsches Statistisches Zentralblatt*, **26**, 137–146.

[7] von Schelling, H., (1954). Coupon Collecting for Unequal Probabilities, Amer. Math. Monthly, **61**, no. 5, 306–311.

[8] Szego, G., (1924). Über eine Eigenschaft der Exponentialreihe, *Sitzungber. Berl. Ges.*, **23**, 50–64. Also in Askey, R. (editor), (1982). *Gabor Szegö: Collected Papers—Volume I (1915-1927).* Birkhauser, Boston.

[9] Wilf, H. (2006). *Generatingfunctionology*, third edition, A K Peters, Wellesley, Massachusetts.

[10] Wong, R. (2001). *Asymptotic Approximations of Integrals*, Classics in Applied Mathematics. SIAM, Philadelphia.