

Notes on the estimation of the asymptotics of the moments for the m collector's problem

Aristides V. Doumas¹

Abstract

The general collector's problem describes a process in which N distinct coupons are placed in an urn and randomly selected one at a time (with replacement) until at least m of all N existing different types of coupons have been selected. Let $T_m(N)$ the random variable denoting the number of trials needed for this goal. We briefly present the leading asymptotics of the (rising) moments of $T_m(N)$ as $N \rightarrow \infty$ for large classes of coupon probabilities. It is proved that the expectation of $T_m(N)$ becomes minimum when the coupons are uniformly distributed. Moreover, a theorem on the asymptotic estimates of the rising moments of $T_m(N)$ by comparison with known sequences of coupon probabilities is proved.

Mathematics Subject Classification: 78M05; 60F99; 41A60

Keywords: Urn problems; coupon collector's problem; double Dixie cup problem; rising moments; Zipf law; Schur functions

¹ Department of Mathematics, National Technical University of Athens, Zografou Campus, 157 80, Athens, Greece. E-mail: aris.doumas@hotmail.com

1 Introduction

We consider the following classical urn problem. Suppose that N distinct types of balls are placed in an urn from which balls are being collected independently with replacement, each one with probability p_j , $j = 1, 2, \dots, N$. Let $T_m(N)$ be the number of trials needed until each ball has been collected m times, where m is a fixed positive integer. This process is, sometimes, called double dixie cup problem, while for the particular case where $m = 1$ is the so-called coupon collector's problem. The problem for the case $m = 1$ has a long history (its origin can be traced back to De Moivre's treatise *De Mensura Sortis* of 1712 and Laplace's pioneering work *Theorie Analytique de Probabilites* of 1812), and its applications lie on several areas of science hence (e.g., biology, linguistics, search algorithms). For general values of m and for $p_j = 1/N$ D. J. Newman and L. Shepp [7] and soonafter, P. Erdős and A. Rényi [5] determined the expectation, as well as the limit distribution of $T_m(N)$. They proved that

$$\lim_{N \rightarrow \infty} P \left\{ \frac{T_m(N) - N \ln N - (m-1)N \ln \ln N + N \ln(m-1)!}{N} \leq y \right\} = e^{-e^{-y}}. \quad (1)$$

For general values of m and for the case of unequal coupon probabilities one may find useful results in [4], where the authors developed techniques of computing the asymptotics of the first and the second moment of $T_m(N)$, the variance, as well as, the limit distribution for large classes of coupon probabilities. Let

$$T_m(N)^{(r)} := T_m(N)(T_m(N) + 1) \cdots (T_m(N) + r - 1), \quad r = 1, 2, \dots \quad (2)$$

i.e., r -th rising moment of $T_m(N)$. In this paper we present leading asymptotics for the rising moments of the random variable $T_m(N)$, for rich classes of probabilities. We prove that $E[T_m(N)^{(r)}]$ becomes minimum when the p_j 's are uniformly distributed by using the Schur - Ostrowski criterion. Finally, a theorem that helps us obtain asymptotic estimates by comparison with sequences of coupon probabilities, for which the asymptotics are known, is presented.

2 The rising moments of $T_m(N)$

Let $\alpha = \{a_j\}_{j=1}^{\infty}$ be a sequence of strictly positive numbers. Then, for each integer $N > 0$, one can create a probability measure $\pi_N = \{p_1, \dots, p_N\}$ on the

set of types $\{1, \dots, N\}$ by taking

$$p_j = \frac{a_j}{A_N}, \quad \text{where} \quad A_N = \sum_{j=1}^N a_j. \quad (3)$$

By a Poissonization technique it is not hard to get explicit formulae for the moments and the moment generating function of $T_m(N)$ (see, [4]):

$$E [T_m(N)^{(r)}] = r \int_0^\infty \left\{ 1 - \prod_{j=1}^N [1 - S_m(p_j t) e^{-p_j t}] \right\} t^{r-1} dt \quad (4)$$

$$G(z) := E [z^{-T_m(N)}] = 1 - (z-1) \int_0^\infty \left\{ 1 - \prod_{j=1}^N [1 - S_m(p_j t) e^{-p_j t}] \right\} e^{-(z-1)t} dt, \quad (5)$$

for $\Re(z) > 1$, $r = 1, 2, \dots$, and $S_m(y)$ denotes the m -th partial sum of e^y , namely

$$S_m(y) := 1 + y + \frac{y^2}{2!} + \dots + \frac{y^{m-1}}{(m-1)!} = \sum_{l=0}^{m-1} \frac{y^l}{l!}. \quad (6)$$

We introduce the notation

$$E_m(N; \alpha; r) := r \int_0^\infty \left[1 - \prod_{j=1}^N \left(1 - e^{-a_j t} S_m(a_j t) \right) \right] t^{r-1} dt. \quad (7)$$

For a sequence $\alpha = \{a_j\}_{j=1}^\infty$ and a number $s > 0$ we set $s\alpha = \{sa_j\}_{j=1}^\infty$. Hence,

$$E [T_m(N)^{(r)}] = A_N^r E_m(N; \alpha; r). \quad (8)$$

Under (8) the problem of estimating $E [T_m(N)^{(r)}]$ can be treated as two separate problems, namely estimating A_N^r and estimating $E_m(N; \alpha; r)$, (see (7)). The estimation of A_N^r can be considered an external matter which can be handled by existing powerful methods, such as the Euler-Maclaurin sum formula, the Laplace method for sums (see, e.g., [1]), or even summation by parts. Let

$$L_m(N; \alpha; r) := \lim_N E_m(N; \alpha; r) = r \int_0^\infty \left[1 - \prod_{j=1}^\infty \left(1 - e^{-a_j t} S_m(a_j t) \right) \right] t^{r-1} dt. \quad (9)$$

Theorem 2.1 *For any fixed positive integers m and r , $E [T_m(N)^{(r)}]$ becomes minimum when all p_j 's are equal.*

Proof. To prove the theorem it suffices to show that, for a fixed $t > 0$, the maximum of the quantity

$$\prod_{j=1}^N [1 - e^{-p_j t} S_m(p_j t)],$$

subject to the constraints $p_1 + \dots + p_N = 1$, $p_j > 0$, $j = 1, 2, \dots, N$, occurs when all p_j 's are equal. Set $\phi : (0, 1)^N \rightarrow (0, \infty)$,

$$\phi(p_1, \dots, p_N) := \sum_{j=1}^N \ln [1 - e^{-p_j t} S_m(p_j t)]. \quad (10)$$

Clearly, ϕ is symmetric w.r.t. its variables. Now, if for all $1 \leq i \neq j \leq N$,

$$(p_i - p_j) \left(\frac{\partial \phi(p_1, p_2, \dots, p_N)}{\partial p_i} - \frac{\partial \phi(p_1, p_2, \dots, p_N)}{\partial p_j} \right) \leq 0, \quad (11)$$

then, ϕ will be a Schur-concave function (see, [6], page 84, theorem A.4) and will attain its maximum when all p_j 's are equal (see, [6], page 413). We have

$$\frac{\partial \phi(p_1, p_2, \dots, p_N)}{\partial p_i} = \frac{t}{(m-1)!} \cdot \frac{e^{-p_i t} (tp_i)^{m-1}}{1 - e^{-p_i t} S_m(p_i t)}.$$

It suffices to obtain that the function $f(\cdot)$ is decreasing, where

$$f(x) := \frac{e^{-x} x^{m-1}}{1 - e^{-x} S_m(x)}, \quad x > 0.$$

Observing that

$$(e^{-x} S_m(x))' = -\frac{e^{-x} x^{m-1}}{(m-1)!},$$

we have

$$f'(x) = \frac{e^{-x} x^{m-2}}{[1 - e^{-x} S_m(x)]^2} g(x), \quad (12)$$

where

$$g(x) := (m-1-x) [1 - e^{-x} S_m(x)] - \frac{e^{-x} x^m}{(m-1)!}. \quad (13)$$

Notice that $g(x)$ extends to a smooth function on \mathbb{R} . In particular $g(0) = 0$. If $m = 1$, then $g(x) = -x$ and (13) implies that $f'(x) < 0$ for all $x > 0$. For $m \geq 2$ we have

$$g'(x) = -1 + e^{-x} S_m(x) - \frac{e^{-x} x^{m-1}}{(m-1)!}, \quad g'(0) = 0,$$

and

$$g''(x) = -(m-1) \frac{e^{-x} x^{m-2}}{(m-1)!} < 0, \quad x > 0.$$

Thus $g(x) < 0$ for all $x > 0$. Therefore, $f'(x) < 0$ for all $x > 0$ and the proof is completed.

The following theorem is related to our recent work [4] and the proof is omitted.

Theorem 2.2 $L_m(N; \alpha; r) < \infty$ simultaneously for all positive (fixed) integers m and r , if and only if there exist a $\xi \in (0, 1)$ such that

$$\sum_{j=1}^{\infty} \xi^{a_j} < \infty.$$

If $L_m(N; \alpha; r) < \infty$, then for all positive integers m and r we have

$$E [T_m(N)^{(r)}] = A_N^r L_m(N; \alpha; r) [1 + o(1)] \text{ as } N \rightarrow \infty.$$

Examples of this case are the *positive power law*, namely $\alpha = \{j^p\}_{j=1}^{\infty}$, where $p > 0$. In particular, when $p = 1$ we have the so-called *linear* case. Also, the families of sequences $\kappa = \{e^{qj}\}_{j=1}^{\infty}$ and where $q > 0$ fall in this case. Notice that the sequences $\beta = \{e^{-qj}\}_{j=1}^{\infty}$ produce the same coupon probabilities with κ , hence they are covered too.

For the challenging case where $L_m(N; \alpha; r) = \infty$ for some fixed positive integer r (and for any fixed m) we write a_j in the form

$$a_j = f(j)^{-1} \tag{14}$$

where

$$f(x) > 0 \quad \text{and} \quad f'(x) > 0, \tag{15}$$

and we will discuss our problem for large classes of distributions. In particular, we will cover the cases where $f(\cdot)$ belongs to the class of positive and strictly increasing $C^3(0, \infty)$ functions, which *grow to ∞ (as $x \rightarrow \infty$) slower than exponentials, but faster than powers of logarithms*. We assume that $f(x)$

possesses three derivatives satisfying the following conditions as $x \rightarrow \infty$:

$$\begin{aligned} \text{(i)} \quad & f(x) \rightarrow \infty, & \text{(ii)} \quad & \frac{f'(x)}{f(x)} \rightarrow 0, \\ \text{(iii)} \quad & \frac{f''(x)/f'(x)}{f'(x)/f(x)} = O(1), & \text{(iv)} \quad & \frac{f'''(x) f(x)^2}{f'(x)^3} = O(1). \end{aligned} \quad (16)$$

These conditions are satisfied by a variety of commonly used functions. For example,

$$f(x) = x^p(\ln x)^q, \quad p > 0, \quad q \in \mathbb{R}, \quad f(x) = \exp(x^r), \quad 0 < r < 1,$$

or various convex combinations of products of such functions. An important example falling in this case is the well known *generalized Zipf law*, namely $f(x) = x^p$, where $p > 0$. Zipf's law has attracted the interest of scientists of several areas of science, such as linguistics, biology, etc.

With similar arguments as in [4] one has the following theorem for the rising moments of the random variable $T_m(N)$.

Theorem 2.3 *If $\alpha = \{1/f(j)\}_{j=1}^\infty$, where $f(\cdot)$ satisfies (15) and (16), then as $N \rightarrow \infty$*

$$E \left[T_N^{(r)} \right] \sim \frac{1}{\min_{1 \leq j \leq N} \{p_j\}^r} \ln \left(\frac{f(N)}{f'(N)} \right)^r. \quad (17)$$

3 Asymptotic estimates for the rising moments of T_N by comparison with known sequences

Here we will present a theorem that helps us obtain asymptotic estimates by comparison with sequences α for which the asymptotic estimates of $E_m(N; \alpha; r)$ are known (for instance, via Theorem 2.3). First, we recall the following notation. Suppose that $\{s_j\}_{j=1}^\infty$ and $\{t_j\}_{j=1}^\infty$ are two sequences of nonnegative terms. The symbol $s_j \asymp t_j$ means that there are two constants $C_1 > C_2 > 0$ and an integer $j_0 > 0$ such that

$$C_2 t_j \leq s_j \leq C_1 t_j, \quad \text{for all } j \geq j_0, \quad (18)$$

i.e. $s_j = O(t_j)$ and $t_j = O(s_j)$.

Theorem 3.1 Let $\alpha = \{a_j\}_{j=1}^{\infty}$ and $\beta = \{b_j\}_{j=1}^{\infty}$ be sequences of strictly positive terms such that $\lim_N E_m(N; \alpha; r) = \lim_N E_m(N; \beta; r) = \infty$.

- (i) If there exists an j_0 such that $a_j = b_j$, for all $j \geq j_0$, then $E_m(N; \alpha; r) - E_m(N; \beta; r)$ is bounded,
(ii) if $a_j = O(b_j)$, then $E_m(N; \beta; r) = O(E_m(N; \alpha; r))$ as $N \rightarrow \infty$,
(iii) if $a_j = o(b_j)$, then $E_m(N; \beta; r) = o(E_m(N; \alpha; r))$ as $N \rightarrow \infty$,
(iv) if $a_j \asymp b_j$, then $E_m(N; \beta; r) \asymp E_m(N; \alpha; r)$ as $N \rightarrow \infty$,
(v) if $a_j \sim b_j$, then $E_m(N; \beta; r) \sim E_m(N; \alpha; r)$ as $N \rightarrow \infty$.

Proof. Case (i) follows easily from (7):

$$\begin{aligned}
& |E_m(N; \alpha; r) - E_m(N; \beta; r)| \\
&= r \left| \int_0^{\infty} \prod_{j=j_0}^N (1 - S_m(a_j t) e^{-a_j t}) \left[\prod_{j=1}^{j_0-1} (1 - S_m(a_j t) e^{-a_j t}) - \prod_{j=1}^{j_0-1} (1 - S_m(b_j t) e^{-b_j t}) \right] t^{r-1} dt \right| \\
&\leq r \int_0^{\infty} \left| \left[\prod_{j=1}^{j_0-1} (1 - S_m(a_j t) e^{-a_j t}) - \prod_{j=1}^{j_0-1} (1 - S_m(b_j t) e^{-b_j t}) \right] \right| t^{r-1} dt \\
&= r \int_0^{\infty} \left| \sum_{J \subset \{1, \dots, j_0-1\}} (-1)^{|J|} \left\{ \exp \left(-t \sum_{j \in J} a_j \right) \prod_{j \in J} S_m(a_j t) \right. \right. \\
&\quad \left. \left. - \exp \left(-t \sum_{j \in J} b_j \right) \prod_{j \in J} S_m(b_j t) \right\} t^{r-1} \right| dt < \infty,
\end{aligned}$$

where we have used the formula

$$\prod_{j=1}^N (1 - S_m(p_j t) e^{-p_j t}) = \sum_{J \subset \{1, \dots, N\}} (-1)^{|J|} \exp \left(-t \sum_{j \in J} p_j \right) \prod_{j \in J} S_m(p_j t). \quad (19)$$

Notice that the sum extends over all 2^{j-1} subsets J of $\{1, \dots, j-1\}$, while $|J|$ denotes the cardinality of J .

(ii) Since $a_j = O(b_j)$, there is a positive constant M and an integer j_0 , such that $a_j \leq M b_j$, for all $j \geq j_0$. By part (i) of the theorem we have

$$|E_m(N; M\beta; r) - E_m(N; \alpha; r)| \leq C, \quad \text{for some positive constant } C \text{ as } N \rightarrow \infty.$$

Next observe that (7) implies

$$E_m(N; s\alpha; r) = s^{-r} E_m(N; \alpha; r). \quad (20)$$

Using (20) we get

$$\left| \frac{1}{M^r} E_m(N; \beta; r) - E_m(N; \alpha; r) \right| \leq C, \quad \text{i.e.} \quad E_m(N; \beta; r) \leq M^r E_m(N; \alpha; r) + CM^r,$$

and the result follows immediately from the definition of the O notation.

(iii) Fix an $\epsilon > 0$. Then $a_j \leq \epsilon b_j$, for all $j \geq j_0(\epsilon)$. Thus, by part (i) there is an $M = M(\epsilon)$ such that

$$E_m(N; \epsilon\beta; r) - E_m(N; \alpha; r) \leq M.$$

By invoking (20) we get

$$\frac{1}{\epsilon^r} E_m(N; \beta; r) \leq E_m(N; \alpha; r) + M, \quad \text{for all } N \geq N_0(\epsilon).$$

If we divide by $E_m(N; \alpha; r)$ and then let $N \rightarrow \infty$, we obtain (iii), since ϵ is arbitrary and $\lim_N E_m(N; \alpha; r) = \infty$.

(iv) Since $a_j \asymp b_j$, then from (18) we have $a_j = O(b_j)$ and $b_j = O(a_j)$. Using part (ii) we get as $N \rightarrow \infty$, $E_m(N; \beta; r) = O(E_m(N; \alpha; r))$ and $E_m(N; \alpha; r) = O(E_m(N; \beta; r))$, the result follows again from (18).

To prove (v) we first fix an $\epsilon > 0$. Then $(1 - \epsilon)b_j \leq a_j \leq (1 + \epsilon)b_j$, for all $j \geq j_0(\epsilon)$. Thus, by case (i) and (20) there is an $M = M(\epsilon)$ such that

$$\left(\frac{1}{1 + \epsilon} \right)^r E_m(N; \beta; r) - M \leq E_m(N; \alpha; r) \leq \left(\frac{1}{1 - \epsilon} \right)^r E_m(N; \beta; r) + M,$$

for all $N \geq N_0(\epsilon)$. If we divide by $E_m(N; \beta; r)$ and then let $N \rightarrow \infty$, we obtain (v) since ϵ is arbitrary and $\lim_N E_m(N; \beta; r) = \infty$.

References

- [1] C.M. Bender and S.A. Orszag, *Advanced Mathematical Methods for Scientists and Engineers I: Asymptotic Methods and Perturbation Theory*, Springer-Verlag, New York, 1999.
- [2] A.V. Dumas and V.G. Papanicolaou, The Coupon Collector's Problem Revisited: Asymptotics of the Variance, *Adv. Appl. Prob.*, **44**(1), (2012), 166–195.

- [3] A.V. Doumas and V.G. Papanicolaou, Asymptotics of the rising moments for the Coupon Collector's Problem, *Electron. J. Probab.*, **18**, (Article no. 41), (2012), 1-15, DOI: 10.1214/EJP.v18-1746.
- [4] A.V. Doumas and V.G. Papanicolaou, The Coupon Collector's Problem Revisited: Generalizing the Double Dixie Cup Problem of Newman and Shepp, *ESAIM: Prob. and Stat.*, Forthcoming article, DOI: <http://dx.doi.org/10.1051/ps/2016016>.
- [5] P. Erdős and A. Rényi, On a classical problem of probability theory, *Magyar. Tud. Akad. Mat. Kutató Int. Közl.*, **6**, (1961), 215-220.
- [6] A. W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and its Applications*, Academic Press, New York, 1979.
- [7] D.J. Newman and L. Shepp, The double Dixie cup problem, *Amer. Math. Monthly*, **67**, (1960), 58-61. MR0120672