Journal of Applied Mathematics & Bioinformatics, Vol.10, No.1, 2020, 1-9
ISSN: 1792-6602 (print), 1792-6939 (online)
Scientific Press International Limited, 2020

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

Aristides V. Doumas[1]

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 $N \to \infty$ 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; Schur - Ostrowski criterion

[1] Department of Mathematics, National Technical University of Athens, Zografou Campus, 157 80, Athens, Greece.
E-mail: aris.doumas@hotmail.com; adou@math.ntua.gr

Article Info: *Received* : July 10, 2019. *Revised* : September 3, 2019.
Published online : January 5, 2020

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, $j = 1, 2, \cdots, N$. 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 [8] and soonafter, P. Erdős and A. Rényi [6] determined the expectation, as well as the limit distribution of $T_m(N)$. They proved that

$$\lim_{N\to\infty} P\left\{\frac{T_m(N) - N\ln N - (m-1)N\ln\ln N + N\ln(m-1)!}{N} \leq y\right\} = e^{-e^{-y}}. \tag{1.1}$$

For general values of m and for the case of unequal coupon probabilities one nay 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

$$T_m(N)^{(r)} := T_m(N)(T_m(N)+1)\cdots(T_m(N)+r-1), \qquad r = 1, 2, \ldots \tag{1.2}$$

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 $\pi_N = \{p_1, ..., p_N\}$ on the set of types $\{1, ..., N\}$ by taking

$$p_j = \frac{a_j}{A_N}, \qquad \text{where} \quad A_N = \sum_{j=1}^{N} a_j. \tag{2.1}$$

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]):

$$E\left[T_m(N)^{(r)}\right] = r \int_0^\infty \left\{1 - \prod_{j=1}^{N} \left[1 - S_m(p_j t) e^{-p_j t}\right]\right\} t^{r-1} dt \tag{2.2}$$

$$\begin{aligned} G(z) \;&:=\; E\left[z^{-T_m(N)}\right] \qquad (2.3)\\ &=\; 1 - (z-1) \int_0^\infty \left\{1 - \prod_{j=1}^{N} \left[1 - S_m(p_j t) e^{-p_j t}\right]\right\} e^{-(z-1)t} dt, \end{aligned}$$

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

$$S_m(y) := 1 + t + \frac{y^2}{2!} + \cdots + \frac{y^{m-1}}{(m-1)!} = \sum_{l=0}^{m-1} \frac{y^l}{l!}\,. \tag{2.4}$$

We introduce the notation

$$E_m(N; \alpha;\, r) : = r \int_0^\infty \left[1 - \prod_{j=1}^{N} \left(1 - e^{-a_j t}\, S_m\left(a_j t\right)\right)\right] t^{r-1} dt. \tag{2.5}$$

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

$$E\left[T_m(N)^{(r)}\right] = A_N^r\, E_m(N; \alpha;\, r). \tag{2.6}$$

Under (2.6) the problem of estimating $E\left[T_m(N)^{(r)}\right]$ can be treated as two separate problems, namely estimating A_N^r and estimating $E_m(N; \alpha;\, r$, (see (2.5)). 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

$$\begin{aligned} L_m(N; \alpha; r) \;&:=\; \lim_N E_m(N; \alpha;\, r) \qquad (2.7)\\ &=\; r \int_0^\infty \left[1 - \prod_{j=1}^{\infty} \left(1 - e^{-a_j t}\, S_m\left(a_j t\right)\right)\right] t^{r-1} dt. \end{aligned}$$

Theorem 2.1. *For any fixed positive integers m and r, $E\left[T_m(N)^{(r)}\right]$ 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

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

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

$$\phi(p_1 \cdots, p_N) := \sum_{j=1}^{N} \ln\left[1 - e^{-p_j t}\, S_m\left(p_j t\right)\right]. \tag{2.8}$$

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

$$\left(p_i - p_j\right)\left(\frac{\partial \phi\left(p_1, p_2, \cdots, p_N\right)}{\partial p_i} - \frac{\partial \phi\left(p_1, p_2, \cdots, p_N\right)}{\partial p_j}\right) \le 0, \tag{2.9}$$

then, ϕ will be a Schur–concave function (see, [7], page 84, theorem A.4) and will attain its maximum when all p_j's are equal (see, [7], page 413). We have

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

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

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

Observing that

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

we have

$$f'(x) = \frac{e^{-x} x^{m-2}}{\left[1 - e^{-x} S_m(x)\right]^2}\, g(x), \tag{2.10}$$

where

$$g(x) := (m - 1 - x)\left[1 - e^{-x} S_m(x)\right] - \frac{e^{-x} x^m}{(m-1)!}. \tag{2.11}$$

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 (2.11) implies that $f'(x) < 0$ for all $x > 0$. For $m \ge 2$ we have

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

and

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

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

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

$$E\left[T_m(N)^{(r)}\right] = A_N^r\, L_m(N;\alpha;r)\,[1+o(1)] \quad \textit{as } N \to \infty.$$

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

$$a_j = f(j)^{-1} \tag{2.12}$$

where

$$f(x) > 0 \qquad \text{and} \qquad f'(x) > 0, \tag{2.13}$$

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 $x \to \infty$) slower* *than exponentials, but faster than powers of logarithms*. We assume that $f(x)$ possesses three derivatives satisfying the following conditions as $x \to \infty$:

$$\text{(i) } f(x) \to \infty, \qquad \text{(ii) } \frac{f'(x)}{f(x)} \to 0,$$
$$\text{(iii) } \frac{f''(x)/f'(x)}{f'(x)/f(x)} = O(1), \qquad \text{(iv) } \frac{f'''(x)\, f(x)^2}{f'(x)^3} = O(1). \tag{2.14}$$

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

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

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 (2.13) and (2.14), then as* $N \to \infty$

$$E\left[T_N^{(r)}\right] \sim \frac{1}{\min_{1 \le j \le N}\{p_j\}^r} \ln\left(\frac{f(N)}{f'(N)}\right)^r. \tag{2.15}$$

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

$$C_2 t_j \le s_j \le C_1 t_j, \qquad \text{for all } j \ge j_0, \tag{3.1}$$

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* $N \to \infty$,
(iii) if $a_j = o(b_j)$, *then* $E_m(N;\beta;r) = o\,(E_m(N;\alpha;r))$ *as* $N \to \infty$,
(iv) if $a_j \asymp b_j$, *then* $E_m(N;\beta;r) \asymp E_m(N;\alpha;r)$ *as* $N \to \infty$,
(v) if $a_j \sim b_j$, *then* $E_m(N;\beta;r) \sim E_m(N;\alpha;r)$ *as* $N \to \infty$.

Proof. Case (i) follows easily from (2.5):

$$|E_m(N;\alpha;\, r) - E_m(N;\beta;\, r)| =$$

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

where we have used the formula

$$\prod_{j=1}^{N}\left(1 - S_m(p_jt)e^{-p_jt}\right) = \sum_{J\subset\{1,\ldots,N\}}(-1)^{|J|}\exp\left(-t\sum_{j\in J}p_j\right)\prod_{j\in J}S_m(p_jt). \quad (3.2)$$

Notice that the sum extends over all 2^{j-1} subsets J of $\{1,\ldots,j-1\}$, 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 $a_j \leq Mb_j$, for all $j \geq j_0$. By part (i) of the theorem we have

$$|E_m(N;M\beta;r) - E_m(N;\alpha;r)| \leq C,$$

for some positive constant C as $N \to \infty$.
Next observe that (2.5) implies

$$E_m(N; s\alpha;\, r) = s^{-r} E_m(N; \alpha;\, r). \tag{3.3}$$

Using (3.3) we get

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

i.e.

$$E_m(N; \beta; r) \leq M^r E_m(N; \alpha; r) + C M^r,$$

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 (3.3) we get

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

If we divide by $E_m(N; \alpha; r)$ and then let $N \to \infty$, 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 (3.1) we have $a_j = O(b_j)$ and $b_j = O(a_j)$. Using part (ii) we get as $N \to \infty$, $E_m(N; \beta; r) = O\left(E_m(N; \alpha; r)\right)$ and $E_m(N; \alpha; r) = O\left(E_m(N; \beta; r)\right)$, the result follows again from (3.1).

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 (3.3) 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 $N \to \infty$, we obtain (v) since ϵ is arbitrary and $\lim_N E_m(N; \beta; r) = \infty$. $\square$

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. Doumas 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**(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: Probability and Statistics*, **20**, (2016), 367–399.

[5] A.V. Doumas and V.G. Papanicolaou, Sampling from a Mixture of Different Groups of Coupons, *arXiv:1709.04500 [math.PR]*, https://arxiv.org/abs/1709.04500.

[6] P. Erdős and A. Rényi, On a classical problem of probability theory, *Magyar. Tud. Akad. Mat. Kutató Int. Közl.*, **6**, (1961), 215–220.

[7] A. W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and its Applications*, Academic Press, New York, 1979.

[8] D.J. Newman and L. Shepp, The double Dixie cup problem, *Amer. Math. Monthly*, **67**, (1960), 58–61.