J. Appl. Prob. **52,** 405–418 (2015)
Printed in England

NEW RESULTS ON A GENERALIZED COUPON COLLECTOR PROBLEM USING MARKOV CHAINS

EMMANUELLE ANCEAUME,* *CNRS*
YANN BUSNEL,** *Université de Nantes*
BRUNO SERICOLA,*** *INRIA*

Abstract

In this paper we study a generalized coupon collector problem, which consists of determining the distribution and the moments of the time needed to collect a given number of distinct coupons that are drawn from a set of coupons with an arbitrary probability distribution. We suppose that a special coupon called the null coupon can be drawn but never belongs to any collection. In this context, we obtain expressions for the distribution and the moments of this time. We also prove that the almost-uniform distribution, for which all the nonnull coupons have the same drawing probability, is the distribution which minimizes the expected time to obtain a fixed subset of distinct coupons. This optimization result is extended to the complementary distribution of the time needed to obtain the full collection, proving by the way this well-known conjecture. Finally, we propose a new conjecture which expresses the fact that the almost-uniform distribution should minimize the complementary distribution of the time needed to obtain any fixed number of distinct coupons.

Keywords: Coupon collector problem; Minimization; Markov chain

2010 Mathematics Subject Classification: Primary 60C05
Secondary 60J05

1. Introduction

The coupon collector problem is an old problem which consists of evaluating the time needed to obtain a collection of different objects drawn randomly using a given probability distribution. This problem has given rise to a lot of attention from researchers in various fields since it has applications in many scientific domains including computer science and optimization; see [1] for several engineering examples.

More formally, consider a set of n coupons which are drawn randomly one by one, with replacement, coupon i being drawn with probability p_i. The classical coupon collector problem is to determine the expectation or the distribution of the number of coupons that need to be drawn from the set of n coupons to obtain the full collection of n coupons. A large number of papers have been devoted to the analysis of asymptotics and limit distributions of this distribution when n tends to infinity; see [3] or [7] and the references therein. In [2], the authors obtain new formulas concerning this distribution and they also provide simulation techniques in order to compute it as well as providing analytic bounds. The asymptotics of the rising moments were studied in [4].

Received 28 February 2014; revision received 10 June 2014.
* Postal address: CNRS, Campus de Beaulieu, 35042 Rennes Cedex, France.
** Postal address: Université de Nantes, 2 rue de la Houssinière, 44322 Nantes Cedex 03, France.
*** Postal address: INRIA, Campus de Beaulieu, 35042 Rennes Cedex, France. Email address: bruno.sericola@inria.fr

In this paper we consider several generalizations of this problem. A first generalization is the analysis, for $c \leq n$, of the number $T_{c,n}$ of coupons that need to be drawn, with replacement, to collect c different coupons from the set $\{1, 2, \ldots, n\}$. With this notation, the number of coupons that need to be drawn from this set to obtain the full collection is $T_{n,n}$. If a coupon is drawn at each discrete time $1, 2, \ldots$ then $T_{c,n}$ is the time needed to obtain c different coupons, this is also called the waiting time to obtain c different coupons. This problem was considered in [8] for the case where the drawing probability distribution is uniform.

In a second generalization, we assume that $\boldsymbol{p} = (p_1, \ldots, p_n)$ is not necessarily a probability distribution, i.e. we suppose that $\sum_{i=1}^{n} p_i \leq 1$ and we define $p_0 = 1 - \sum_{i=1}^{n} p_i$. This means that there is a null coupon, denoted by 0, which is drawn with probability p_0, but which does not belong to the collection. In this context, the problem is to determine the distribution of the number $T_{c,n}$ of coupons that need to be drawn from set $\{0, 1, \ldots, n\}$, with replacement, until we first obtain a collection composed of c different coupons, $1 \leq c \leq n$, among $\{1, \ldots, n\}$. This work is motivated by the analysis of streaming algorithms in network monitoring applications, presented in Section 7.

In Section 2 the distribution of $T_{c,n}$ is obtained using Markov chains. Moreover, we show that this distribution leads to new combinatorial identities. This result is used to derive an expression of $T_{c,n}(\boldsymbol{v})$ when the drawing distribution is the almost-uniform distribution denoted by $\boldsymbol{v}$ and defined by $\boldsymbol{v} = (v_1, \ldots, v_n)$ with $v_i = (1 - v_0)/n$, where $v_0 = 1 - \sum_{i=1}^{n} v_i$. Expressions for the moments of $T_{c,n}(\boldsymbol{p})$ are presented in Section 3, where we show that the limit of $\mathbb{E}\{T_{c,n}(\boldsymbol{p})\}$ is equal to c when n tends to ∞. In Section 4 we show that the almost-uniform distribution $\boldsymbol{v}$ and the uniform distribution $\boldsymbol{u}$ minimize the expected value $\mathbb{E}\{T_{c,n}(\boldsymbol{p})\}$. In Section 5 we prove that the tail distribution of $T_{n,n}$ is minimized over all the $p_1, \ldots, p_n$ by the almost-uniform distribution and by the uniform distribution. This result was expressed as a conjecture in the case where $p_0 = 0$, i.e. when $\sum_{i=1}^{n} p_i = 1$, in several papers; see, for example, [1] from which the idea of the proof originates. In Section 6 we propose a new conjecture which consists of showing that the distributions $\boldsymbol{v}$ and $\boldsymbol{u}$ minimize the tail distribution of $T_{c,n}(\boldsymbol{p})$. This conjecture is motivated by the fact that it is true for $c = 1$ and $c = n$ as shown in Section 5, and we show that it is also true for $c = 2$. It is, moreover, true for the expected value $\mathbb{E}\{T_{c,n}(\boldsymbol{p})\}$ as shown in Section 4.

2. Distribution of $T_{c,n}$

Recall that $T_{c,n}$ is the number of coupons that need to be drawn from the set $\{0, 1, 2, \ldots, n\}$, with replacement, until we first obtain a collection with c different coupons, $1 \leq c \leq n$, among $\{1, \ldots, n\}$, where coupon i is drawn with probability p_i, $i = 0, 1, \ldots, n$. To obtain the distribution of $T_{c,n}$, we consider a discrete-time Markov chain $X = \{X_m, m \geq 0\}$ that represents the collection obtained after having drawn m coupons. The state space of X is $S_n = \{J \subseteq \{1, \ldots, n\}\}$ and its transition probability matrix, denoted by Q is given, for every $J, H \in S_n$, by

$$Q_{J,H} = \begin{cases} p_\ell & \text{if } H \setminus J = \{\ell\}, \\ p_0 + P_J & \text{if } J = H, \\ 0 & \text{otherwise,} \end{cases}$$

where for every $J \in S_n$, P_J is given by $P_J = \sum_{j \in J} p_j$, with $P_\varnothing = 0$. It is easily checked that Markov chain X is acyclic, i.e. it has no cycle of length greater than 1, and that all the states are transient, except state $\{1, \ldots, n\}$ which is absorbing. We introduce the partition

https://doi.org/10.1239/jap/1437658606 Published online by Cambridge University Press

$(S_{0,n}, S_{1,n}, \ldots, S_{n,n})$ of S_n, where $S_{i,n}$ is defined for $i = 0, \ldots, n$, by

$$S_{i,n} = \{J \subseteq \{1, \ldots, n\} \mid |J| = i\}. \tag{1}$$

Note that we have $S_{0,n} = \{\varnothing\}$, $|S_n| = 2^n$, and $|S_{i,n}| = \binom{n}{i}$. Assuming that $X_0 = \varnothing$ with probability 1, the random variable $T_{c,n}$ can then be defined for every $c = 1, \ldots, n$, by

$$T_{c,n} = \inf\{m \geq 0 \mid X_m \in S_{c,n}\}.$$

The distribution of $T_{c,n}$ is obtained in Theorem 1 using the Markov property and the following lemma. For every $n \geq 1$, $\ell = 1, \ldots, n$ and $i = 0, \ldots, n$, we define the set $S_{i,n}(\ell)$ by $S_{i,n}(\ell) = \{J \subseteq \{1, \ldots, n\} \setminus \{\ell\} \mid |J| = i\}$.

Lemma 1. *For every $n \geq 1$, $k \geq 0$, and for all positive real numbers $y_1, \ldots, y_n$, for every $i = 1, \ldots, n$, and all real numbers $a \geq 0$, we have*

$$\sum_{\ell=1}^{n} y_\ell \sum_{J \in S_{i-1,n}(\ell)} (a + y_\ell + Y_J)^k = \sum_{J \in S_{i,n}} Y_J (a + Y_J)^k,$$

where $Y_J = \sum_{j \in J} y_j$ and $Y_\varnothing = 0$.

Proof. For $n = 1$, since $S_{0,1}(1) = \varnothing$, the left-hand side is equal to $y_1(a + y_1)^k$ and since $S_{1,1} = \{1\}$, the right-hand side is also equal to $y_1(a + y_1)^k$. Suppose that the result is true for integer $n - 1$ i.e. suppose that

$$\sum_{\ell=1}^{n-1} y_\ell \sum_{J \in S_{i-1,n-1}(\ell)} (a + y_\ell + Y_J)^k = \sum_{J \in S_{i,n-1}} Y_J (a + Y_J)^k.$$

Then

$$\sum_{\ell=1}^{n} y_\ell \sum_{J \in S_{i-1,n}(\ell)} (a+y_\ell+Y_J)^k = \sum_{\ell=1}^{n-1} y_\ell \sum_{J \in S_{i-1,n}(\ell)} (a+y_\ell+Y_J)^k + y_n \sum_{J \in S_{i-1,n}(n)} (a+y_n+Y_J)^k.$$

Since $S_{i-1,n}(n) = S_{i-1,n-1}$, we obtain

$$\sum_{\ell=1}^{n} y_\ell \sum_{J \in S_{i-1,n}(\ell)} (a+y_\ell+Y_J)^k = \sum_{\ell=1}^{n-1} y_\ell \sum_{J \in S_{i-1,n}(\ell)} (a+y_\ell+Y_J)^k + y_n \sum_{J \in S_{i-1,n-1}} (a+y_n+Y_J)^k.$$

For $\ell = 1, \ldots, n-1$, the set $S_{i-1,n}(\ell)$ can be partitioned into two subsets $S'_{i-1,n}(\ell)$ and $S''_{i-1,n}(\ell)$ defined by

$$S'_{i-1,n}(\ell) = \{J \subseteq \{1, \ldots, n\} \setminus \{\ell\} \mid |J| = i - 1 \text{ and } n \in J\}$$

and

$$S''_{i-1,n}(\ell) = \{J \subseteq \{1, \ldots, n\} \setminus \{\ell\} \mid |J| = i - 1 \text{ and } n \notin J\}.$$

https://doi.org/10.1239/jap/1437658606 Published online by Cambridge University Press

Since $S''_{i-1,n}(\ell) = S_{i-1,n-1}(\ell)$, the previous relation becomes

$$\begin{aligned}
&\sum_{\ell=1}^{n} y_\ell \sum_{J\in S_{i-1,n}(\ell)} (a + y_\ell + Y_J)^k \\
&\quad = \sum_{\ell=1}^{n-1} y_\ell \Bigg[\sum_{J\in S_{i-1,n-1}(\ell)} (a + y_\ell + Y_J)^k + \sum_{J\in S'_{i-1,n}(\ell)} (a + y_\ell + Y_J)^k \Bigg] \\
&\qquad + y_n \sum_{J\in S_{i-1,n-1}} (a + y_n + Y_J)^k \\
&\quad = \sum_{\ell=1}^{n-1} y_\ell \sum_{J\in S_{i-1,n-1}(\ell)} (a + y_\ell + Y_J)^k + \sum_{\ell=1}^{n-1} y_\ell \sum_{J\in S_{i-2,n-1}(\ell)} (a + y_n + y_\ell + Y_J)^k \\
&\qquad + y_n \sum_{J\in S_{i-1,n-1}} (a + y_n + Y_J)^k.
\end{aligned}$$

The recurrence hypothesis can be applied for both the first and the second terms. For the second term, the constant a is replaced by the constant $a + y_n$. Thus,

$$\begin{aligned}
&\sum_{\ell=1}^{n} y_\ell \sum_{J\in S_{i-1,n}(\ell)} (a + y_\ell + Y_J)^k \\
&\quad = \sum_{J\in S_{i,n-1}} Y_J (a + Y_J)^k + \sum_{J\in S_{i-1,n-1}} Y_J (a + y_n + Y_J)^k + y_n \sum_{J\in S_{i-1,n-1}} (a + y_n + Y_J)^k \\
&\quad = \sum_{J\in S_{i,n-1}} Y_J (a + Y_J)^k + \sum_{J\in S_{i-1,n-1}} (y_n + Y_J)(a + y_n + Y_J)^k \\
&\quad = \sum_{J\in S_{i,n-1}} Y_J (a + Y_J)^k + \sum_{J\in S'_{i,n}} Y_J (a + Y_J)^k,
\end{aligned}$$

where $S'_{i,n} = \{J \subseteq \{1, \ldots, n\} \mid |J| = i \text{ and } n \in J\}$.

Consider the set $S''_{i,n} = \{J \subseteq \{1, \ldots, n\} \mid |J| = i \text{ and } n \notin J\}$. The sets $S'_{i,n}$ and $S''_{i,n}$ form a partition of $S_{i,n}$ and since $S''_{i,n} = S_{i,n-1}$, we obtain

$$\begin{aligned}
\sum_{\ell=1}^{n} y_\ell \sum_{J\in S_{i-1,n}(\ell)} (a + y_\ell + Y_J)^k &= \sum_{J\in S_{i,n-1}} Y_J (a + Y_J)^k + \sum_{J\in S'_{i,n}} Y_J (a + Y_J)^k \\
&= \sum_{J\in S''_{i,n}} Y_J (a + Y_J)^k + \sum_{J\in S'_{i,n}} Y_J (a + Y_J)^k \\
&= \sum_{J\in S_{i,n}} Y_J (a + Y_J)^k,
\end{aligned}$$

which completes the proof.

In the following we will use the fact that the distribution of $T_{c,n}$ depends on the vector $\boldsymbol{p} = (p_1, \ldots, p_n)$, so we will use the notation $T_{c,n}(\boldsymbol{p})$ instead of $T_{c,n}$, meaning by the way

https://doi.org/10.1239/jap/1437658606 Published online by Cambridge University Press

that vector $\boldsymbol{p}$ is of dimension n. We will also use the notation $p_0 = 1 - \sum_{i=1}^{n} p_i$. Finally, for $\ell = 1, \ldots, n$, the notation $\boldsymbol{p}^{(\ell)}$ will denote the vector $\boldsymbol{p}$ in which the entry p_ℓ has been removed, that is, $\boldsymbol{p}^{(\ell)} = (p_i, 1 \le i \le n, i \ne \ell)$. The dimension of $\boldsymbol{p}^{(\ell)}$, which is $n-1$ here, is not specified but will be clear by the context of its use.

Theorem 1. *For every $n \ge 1$ and $c = 1, \ldots, n$, we have for every $k \ge 0$,*

$$\mathbb{P}\{T_{c,n}(\boldsymbol{p}) > k\} = \sum_{i=0}^{c-1} (-1)^{c-1-i} \binom{n-i-1}{n-c} \sum_{J \in S_{i,n}} (p_0 + P_J)^k. \tag{2}$$

Proof. It holds that (2) is true for $c = 1$ since in this case we have $\mathbb{P}\{T_{1,n}(\boldsymbol{p}) > k\} = p_0^k$. So we now suppose that $n \ge 2$ and $c = 2, \ldots, n$. Since $X_0 = \varnothing$, conditioning on X_1 and using the Markov property (see, for example, [9]) it follows that for $k \ge 1$,

$$\mathbb{P}\{T_{c,n}(\boldsymbol{p}) > k\} = p_0 \mathbb{P}\{T_{c,n}(\boldsymbol{p}) > k-1\} + \sum_{\ell=1}^{n} p_\ell \mathbb{P}\{T_{c-1,n-1}(\boldsymbol{p}^{(\ell)}) > k-1\}. \tag{3}$$

We now proceed by recurrence over k. It holds that (2) is true for $k = 0$ since it is well known that

$$\sum_{i=0}^{c-1} (-1)^{c-1-i} \binom{n-i-1}{n-c} \binom{n}{i} = 1. \tag{4}$$

It holds that (2) is also true for $k = 1$ since, on the one hand, $\mathbb{P}\{T_{c,n}(\boldsymbol{p}) > 1\} = 1$ and on the other hand, using (3), we have

$$\mathbb{P}\{T_{c,n}(\boldsymbol{p}) > 1\} = p_0 \mathbb{P}\{T_{c,n}(\boldsymbol{p}) > 0\} + \sum_{\ell=1}^{n} p_\ell \mathbb{P}\{T_{c-1,n-1}(\boldsymbol{p}^{(\ell)}) > 0\} = p_0 + \sum_{\ell=1}^{n} p_\ell = 1.$$

Suppose now that (2) is true for integer $k-1$, that is, suppose that we have

$$\mathbb{P}\{T_{c,n}(\boldsymbol{p}) > k-1\} = \sum_{i=0}^{c-1} (-1)^{c-1-i} \binom{n-i-1}{n-c} \sum_{J \in S_{i,n}} (p_0 + P_J)^{k-1}.$$

Using (3) and the recurrence relation, we have

$$\begin{aligned}\mathbb{P}\{T_{c,n}(\boldsymbol{p}) > k\} = {} & p_0 \sum_{i=0}^{c-1} (-1)^{c-1-i} \binom{n-i-1}{n-c} \sum_{J \in S_{i,n}} (p_0 + P_J)^{k-1} \\ & + \sum_{\ell=1}^{n} p_\ell \sum_{i=0}^{c-2} (-1)^{c-2-i} \binom{n-i-2}{n-c} \sum_{J \in S_{i,n}(\ell)} (p_0 + p_\ell + P_J)^{k-1}.\end{aligned}$$

Using the change of variable $i := i - 1$ in the second sum, we obtain

$$\begin{aligned}\mathbb{P}\{T_{c,n}(\boldsymbol{p}) > k\} = {} & \sum_{i=0}^{c-1} (-1)^{c-1-i} \binom{n-i-1}{n-c} p_0 \sum_{J \in S_{i,n}} (p_0 + P_J)^{k-1} \\ & + \sum_{i=1}^{c-1} (-1)^{c-1-i} \binom{n-i-1}{n-c} \sum_{\ell=1}^{n} p_\ell \sum_{J \in S_{i-1,n}(\ell)} (p_0 + p_\ell + P_J)^{k-1}\end{aligned}$$

https://doi.org/10.1239/jap/1437658606 Published online by Cambridge University Press

$$= \sum_{i=1}^{c-1} (-1)^{c-1-i} \binom{n-i-1}{n-c} \times \left[p_0 \sum_{J \in S_{i,n}} (p_0 + P_J)^{k-1} + \sum_{\ell=1}^{n} p_\ell \sum_{J \in S_{i-1,n}(\ell)} (p_0 + p_\ell + P_J)^{k-1} \right] + (-1)^{c-1} \binom{n-1}{n-c} p_0^k.$$

From Lemma 1, we have

$$\sum_{\ell=1}^{n} p_\ell \sum_{J \in S_{i-1,n}(\ell)} (p_0 + p_\ell + P_J)^{k-1} = \sum_{J \in S_{i,n}} P_J (p_0 + P_J)^{k-1},$$

that is,

$$\mathbb{P}\{T_{c,n}(\boldsymbol{p}) > k\} = (-1)^{c-1} \binom{n-1}{n-c} p_0^k + \sum_{i=1}^{c-1} (-1)^{c-1-i} \binom{n-i-1}{n-c} \sum_{J \in S_{i,n}} (p_0 + P_J)^k$$
$$= \sum_{i=0}^{c-1} (-1)^{c-1-i} \binom{n-i-1}{n-c} \sum_{J \in S_{i,n}} (p_0 + P_J)^k,$$

which completes the proof.

This theorem also shows, as expected, that the function $\mathbb{P}\{T_{c,n}(\boldsymbol{p}) > k\}$, as a function of $\boldsymbol{p}$, is symmetric, which means that it has the same value for any permutation of the entries of $\boldsymbol{p}$. As a corollary, we obtain the following combinatorial identities.

Corollary 1. *For all $c \geq 1$, $n \geq c$, and $p_1, \ldots, p_n \in (0, 1)$ such that $\sum_{i=1}^{n} p_i = 1$,*

$$\sum_{i=0}^{c-1} (-1)^{c-1-i} \binom{n-i-1}{n-c} \sum_{J \in S_{i,n}} (p_0 + P_J)^{k-1} = 1 \quad \text{for } k = 0, 1, \ldots, c-1.$$

Proof. Use Theorem 1 and the fact that $T_{c,n} \geq c$ with probability 1.

For all $n \geq 1$ and $v_0 \in [0, 1]$, we define the vector $\boldsymbol{v} = (v_1, \ldots, v_n)$ by $v_i = (1 - v_0)/n$. We will refer it to as the almost-uniform distribution. We then have, from (2),

$$\mathbb{P}\{T_{c,n}(\boldsymbol{v}) > k\} = \sum_{i=0}^{c-1} (-1)^{c-1-i} \binom{n-i-1}{n-c} \binom{n}{i} \left(v_0 \left(1 - \frac{i}{n}\right) + \frac{i}{n} \right)^k.$$

We denote by $\boldsymbol{u} = (u_1, \ldots, u_n)$ the uniform distribution defined by $u_i = 1/n$. It is equal to $\boldsymbol{v}$ when $v_0 = 0$. The dimensions of $\boldsymbol{u}$ and $\boldsymbol{v}$ are specified by the context.

3. Moments of $T_{c,n}$

For $r \geq 1$, the rth moment of $T_{c,n}(\boldsymbol{p})$ is defined by

$$\mathbb{E}\{T_{c,n}^r(\boldsymbol{p})\} = \sum_{k=1}^{\infty} k^r \mathbb{P}\{T_{c,n}(\boldsymbol{p}) = k\} = \sum_{\ell=0}^{r-1} \left\langle {r \atop \ell} \right\rangle \sum_{k=0}^{\infty} k^\ell \mathbb{P}\{T_{c,n}(\boldsymbol{p}) > k\}.$$

https://doi.org/10.1239/jap/1437658606 Published online by Cambridge University Press

The first moment of $T_{c,n}(\boldsymbol{p})$ is then obtained by taking $r = 1$, that is,

$$\mathbb{E}\{T_{c,n}(\boldsymbol{p})\} = \sum_{i=0}^{c-1} (-1)^{c-1-i} \binom{n-i-1}{n-c} \sum_{J \in S_{i,n}} \frac{1}{1-(p_0 + P_J)}. \tag{5}$$

The expected value (5) was obtained in [5] in the particular case where $p_0 = 0$. When the drawing probabilities are given by the almost-uniform distribution $\boldsymbol{v}$, we obtain

$$\mathbb{E}\{T_{c,n}(\boldsymbol{v})\} = \frac{1}{1-v_0} \sum_{i=0}^{c-1} (-1)^{c-1-i} \binom{n-i-1}{n-c} \binom{n}{i} \frac{n}{n-i} = \frac{1}{1-v_0} \mathbb{E}\{T_{c,n}(\boldsymbol{u})\}.$$

Using the relation

$$\binom{n}{i} \frac{n}{n-i} = \binom{n}{i} + \binom{n-1}{i-1} \frac{n}{n-i} \mathbf{1}_{\{i \geq 1\}},$$

where $\mathbf{1}_A$ is the indicator function of set A, we obtain

$$\mathbb{E}\{T_{c,n}(\boldsymbol{u})\} = \sum_{i=0}^{c-1} (-1)^{c-1-i} \binom{n-i-1}{n-c} \left[\binom{n}{i} + \binom{n-1}{i-1} \frac{n}{n-i} \mathbf{1}_{\{i \geq 1\}} \right].$$

Using (4) and the change of variable $i := i + 1$, we obtain

$$\mathbb{E}\{T_{c,n}(\boldsymbol{u})\} = 1 + \frac{n}{n-1} \mathbb{E}\{T_{c-1,n-1}(\boldsymbol{u})\}. \tag{6}$$

Note that dimension of the uniform distribution in the left-hand side is equal to n and the one on the right-hand side is equal to $n-1$. Since $\mathbb{E}\{T_{1,n}(\boldsymbol{u})\} = 1$, we obtain

$$\mathbb{E}\{T_{c,n}(\boldsymbol{u})\} = n(H_n - H_{n-c}), \qquad \mathbb{E}\{T_{c,n}(\boldsymbol{v})\} = \frac{n(H_n - H_{n-c})}{1-v_0}, \tag{7}$$

where H_ℓ is the ℓth harmonic number defined by $H_0 = 0$ and $H_\ell = \sum_{i=1}^{\ell} 1/i$ for $\ell \geq 1$. We deduce easily from (6) that for every $c \geq 1$, we have

$$\lim_{n \to \infty} \mathbb{E}\{T_{c,n}(\boldsymbol{u})\} = c, \qquad \lim_{n \to \infty} \mathbb{E}\{T_{c,n}(\boldsymbol{v})\} = \frac{c}{1-v_0}.$$

In the next section we show that when p_0 is fixed the minimum value of $\mathbb{E}\{T_{c,n}(\boldsymbol{p})\}$ is reached when $\boldsymbol{p} = \boldsymbol{v}$, with $v_0 = p_0$.

4. The distribution minimizing $\mathbb{E}\{T_{c,n}(\boldsymbol{p})\}$

The following lemma will be used to prove the next theorem.

Lemma 2. *For $n \geq 1$ and $r_1, \ldots, r_n > 0$ with $\sum_{\ell=1}^{n} r_\ell = 1$, we have $\sum_{\ell=1}^{n} 1/r_\ell \geq n^2$.*

Proof. The result is true for $n = 1$. Suppose it is true for integer $n-1$. We have

$$\sum_{\ell=1}^{n} \frac{1}{r_\ell} = \frac{1}{r_n} + \sum_{\ell=1}^{n-1} \frac{1}{r_\ell} = \frac{1}{r_n} + \frac{1}{1-r_n} \sum_{\ell=1}^{n-1} \frac{1}{h_\ell},$$

where $h_\ell = r_\ell/(1-r_n)$. Since $\sum_{\ell=1}^{n-1} h_\ell = 1$, using the recurrence hypothesis it follows that

$$\sum_{\ell=1}^{n} \frac{1}{r_\ell} \geq \frac{1}{r_n} + \frac{(n-1)^2}{1-r_n} = \frac{(nr_n - 1)^2}{r_n(1-r_n)} + n^2 \geq n^2.$$

https://doi.org/10.1239/jap/1437658606 Published online by Cambridge University Press

Theorem 2. *For every $n \geq 1$, $c = 1, \ldots, n$, and $\boldsymbol{p} = (p_1, \ldots, p_n) \in (0,1)^n$ with $\sum_{i=1}^n p_i \leq 1$, we have $\mathbb{E}\{T_{c,n}(\boldsymbol{p})\} \geq \mathbb{E}\{T_{c,n}(\boldsymbol{v})\} \geq \mathbb{E}\{T_{c,n}(\boldsymbol{u})\}$, where $\boldsymbol{v} = (v_1, \ldots, v_n)$ with $v_i = (1 - p_0)/n$ and $p_0 = 1 - \sum_{i=1}^n p_i$ and where $\boldsymbol{u} = (1/n, \ldots, 1/n)$.*

Proof. The second inequality comes from (7). Defining $v_0 = 1 - \sum_{i=1}^n v_i$, we have $v_0 = p_0$. For $c = 1$ we have, from (5), $\mathbb{E}\{T_{1,n}(\boldsymbol{p})\} = 1/(1-p_0) = 1/(1-v_0) = \mathbb{E}\{T_{1,n}(\boldsymbol{v})\}$. For $c \geq 2$, which implies that $n \geq 2$, summing (3) for $k \geq 1$, we obtain

$$\mathbb{E}\{T_{c,n}(\boldsymbol{p})\} = \frac{1}{1-p_0}\left(1 + \sum_{\ell=1}^{n} p_\ell \mathbb{E}\{T_{c-1,n-1}(\boldsymbol{p}^{(\ell)})\}\right). \tag{8}$$

Suppose that the inequality is true for integer $c-1$, i.e. suppose that for $n \geq c$, and for $\boldsymbol{q} = (q_1, \ldots, q_{n-1}) \in (0,1)^{n-1}$ with $\sum_{i=1}^{n-1} q_i \leq 1$, we have $\mathbb{E}\{T_{c-1,n-1}(\boldsymbol{q})\} \geq \mathbb{E}\{T_{c-1,n-1}(\boldsymbol{v})\}$ with $v_0 = q_0 = 1 - \sum_{i=1}^{n-1} q_i$. Using (7), this implies

$$\mathbb{E}\{T_{c-1,n-1}(\boldsymbol{p}^{(\ell)})\} \geq \frac{(n-1)(H_{n-1} - H_{n-c})}{1 - (p_0 + p_\ell)}.$$

From (8), we obtain

$$\mathbb{E}\{T_{c,n}(\boldsymbol{p})\} \geq \frac{1}{1-p_0}\left(1 + (n-1)(H_{n-1} - H_{n-c}) \sum_{\ell=1}^{n} \frac{p_\ell}{1 - (p_0 + p_\ell)}\right). \tag{9}$$

Observe that for $\ell = 1, \ldots, n$, we have

$$\frac{p_\ell}{1 - (p_0 + p_\ell)} = -1 + \frac{1}{(n-1)r_\ell},$$

where $r_\ell = (1 - (p_0 + p_\ell))/(n-1)(1-p_0)$. These r_ℓ satisfy $r_1, \ldots, r_n > 0$ with $\sum_{\ell=1}^n r_\ell = 1$. From Lemma 2, we obtain

$$\sum_{\ell=1}^{n} \frac{p_\ell}{1 - (p_0 + p_\ell)} = -n + \frac{1}{n-1} \sum_{\ell=1}^{n} \frac{1}{r_\ell} \geq -n + \frac{n^2}{n-1} = \frac{n}{n-1}.$$

Substituting this into (9), we obtain, using (7),

$$\mathbb{E}\{T_{c,n}(\boldsymbol{p})\} \geq \frac{1}{1-p_0}(1 + n(H_{n-1} - H_{n-c})) = \frac{n(H_n - H_{n-c})}{1-p_0} = \mathbb{E}\{T_{c,n}(\boldsymbol{v})\}.$$

5. The distribution minimizing the distribution of $T_{n,n}(\boldsymbol{p})$

For all $n \geq 1$, $i = 0, \ldots, n$, and $k \geq 0$, we denote by $N_i^{(k)}$ the number of coupons of type i collected at instants $1, \ldots, k$. It is well known that the joint distribution of the $N_i^{(k)}$ is a multinomial distribution, i.e. for all $k_0, \ldots, k_n \geq 0$ such that $\sum_{i=0}^n k_i = k$,

$$\mathbb{P}\{N_0^{(k)} = k_0, N_1^{(k)} = k_1, \ldots, N_n^{(k)} = k_n\} = \frac{k!}{k_0!\, k_1! \cdots k_n!} p_0^{k_0} p_1^{k_1} \cdots p_n^{k_n}. \tag{10}$$

Recall that the coupons of type 0 do not belong to the collection. For every $\ell = 1, \ldots, n$, we easily deduce that for every $k \geq 0$ and $k_1, \ldots, k_\ell \geq 0$ such that $\sum_{i=1}^{\ell} k_i \leq k$,

$$\mathbb{P}\{N_1^{(k)} = k_1, \ldots, N_\ell^{(k)} = k_\ell\} = \frac{k!\, p_1^{k_1} \cdots p_\ell^{k_\ell}}{k_1! \cdots k_\ell!\, (k - \sum_{i=1}^{\ell} k_i)!} \left(1 - \sum_{i=1}^{\ell} p_i\right)^{k - \sum_{i=1}^{\ell} k_i}.$$

https://doi.org/10.1239/jap/1437658606 Published online by Cambridge University Press

To prove the next theorem, we recall some basic results on convex functions. Let f be a function defined on an interval I. For all $\alpha \in I$, we introduce the function g_α, defined for all $x \in I \setminus \{\alpha\}$ by $g_\alpha(x) = (f(x) - f(\alpha))/(x - \alpha)$. It is an easy exercise to check that f is convex on interval I if and only if for all $\alpha \in I$, g_α is increasing on $I \setminus \{\alpha\}$. The next result is also known but less popular, so we will provide the proof.

Lemma 3. *Let f be a convex function on an interval I. For every $x, y, z, t \in I$ with $x < y$, $z < t$, we have $(t - y)f(z) + (z - x)f(y) \le (t - y)f(x) + (z - x)f(t)$. If, moreover, we have $t + x = y + z$, we obtain $f(z) + f(y) \le f(x) + f(t)$.*

Proof. We apply twice the fact that g_α is increasing on $I \setminus \{\alpha\}$ for all $\alpha \in I$. Since $z < t$ and $x < y$, we have $g_x(z) \le g_x(t)$ and $g_t(x) \le g_t(y)$. But as $g_x(t) = g_t(x)$ and $g_t(y) = g_y(t)$, we obtain $g_x(z) \le g_x(t) = g_t(x) \le g_t(y) = g_y(t)$, which means that

$$\frac{f(z) - f(x)}{z - x} \le \frac{f(t) - f(y)}{t - y},$$

that is, $(t - y)f(z) + (z - x)f(y) \le (t - y)f(x) + (z - x)f(t)$. The rest of the proof is trivial since $t + x = y + z$ implies that $t - y = z - x > 0$.

Theorem 3. *For all $n \ge 1$ and $\boldsymbol{p} = (p_1, \dots, p_n) \in (0, 1)^n$ with $\sum_{i=1}^n p_i \le 1$, we have for all $k \ge 0$, $\mathbb{P}\{T_{n,n}(\boldsymbol{p}') \le k\} \ge \mathbb{P}\{T_{n,n}(\boldsymbol{p}) \le k\}$, where $\boldsymbol{p}' = (p_1, \dots, p_{n-2}, p'_{n-1}, p'_n)$ with $p'_{n-1} = \lambda p_{n-1} + (1 - \lambda)p_n$ and $p'_n = (1 - \lambda)p_{n-1} + \lambda p_n$ for all $\lambda \in [0, 1]$.*

Proof. If $\lambda = 1$ then we have $\boldsymbol{p}' = \boldsymbol{p}$ so the result is trivial. If $\lambda = 0$ then we have $p'_{n-1} = p_n$ and $p'_n = p_{n-1}$ and the result is also trivial since the function $\mathbb{P}\{T_{n,n}(\boldsymbol{p}) \le k\}$ is a symmetric function of $\boldsymbol{p}$. Thus, we now suppose that $\lambda \in (0, 1)$. For every $n \ge 1$ and $k \ge 0$, we have $\{T_{n,n}(\boldsymbol{p}) \le k\} = \{N_1^{(k)} > 0, \dots, N_n^{(k)} > 0\}$. Thus, we obtain for $k_1, \dots, k_{n-2} > 0$ such that $\sum_{i=1}^{n-2} k_i \le k$ and setting $s = k - \sum_{i=1}^{n-2} k_i$,

$$\begin{aligned}
&\mathbb{P}\{T_{n,n}(\boldsymbol{p}) \le k, N_1^{(k)} = k_1, \dots, N_{n-2}^{(k)} = k_{n-2}\} \\
&\quad = \mathbb{P}\{N_1^{(k)} = k_1, \dots, N_{n-2}^{(k)} = k_{n-2}, N_{n-1}^{(k)} > 0, N_n^{(k)} > 0\} \\
&\quad = \sum_{(u,v,w)\in\mathcal{A}} \mathbb{P}\{N_0^{(k)} = u, N_1^{(k)} = k_1, \dots, N_{n-2}^{(k)} = k_{n-2}, N_{n-1}^{(k)} = v, N_n^{(k)} = w\},
\end{aligned}$$

where $\mathcal{A} = \{(u, v, w) \mid u \ge 0, v > 0, w > 0, u + v + w = s\}$. Using (10) and introducing $q_0 = p_0/(p_0 + p_{n-1} + p_n)$, $q_{n-1} = p_{n-1}/(p_0 + p_{n-1} + p_n)$, and $q_n = p_n/(p_0 + p_{n-1} + p_n)$, we obtain

$$\begin{aligned}
&\mathbb{P}\{T_{n,n}(\boldsymbol{p}) \le k, N_1^{(k)} = k_1, \dots, N_{n-2}^{(k)} = k_{n-2}\} \\
&\quad = \sum_{(u,v,w)\in\mathcal{A}} \frac{k!\, p_0^u p_1^{k_1} \cdots p_{n-2}^{k_{n-2}} p_{n-1}^v p_n^w}{u!\, k_1! \cdots k_{n-2}!\, v!\, w!} = \frac{k!\, p_1^{k_1} \cdots p_{n-2}^{k_{n-2}}}{k_1! \cdots k_{n-2}!} \sum_{(u,v,w)\in\mathcal{A}} \frac{p_0^u p_{n-1}^v p_n^w}{u!\, v!\, w!} \\
&\quad = \frac{k!\, p_1^{k_1} \cdots p_{n-2}^{k_{n-2}} (1 - (p_1 + \cdots + p_{n-2}))^s}{k_1! \cdots k_{n-2}!\, s!} \sum_{(u,v,w)\in\mathcal{A}} \frac{s!}{u!\, v!\, w!} q_0^u q_{n-1}^v q_n^w \\
&\quad = \frac{k!\, p_1^{k_1} \cdots p_{n-2}^{k_{n-2}} (1 - (p_1 + \cdots + p_{n-2}))^s}{k_1! \cdots k_{n-2}!\, s!} (1 - (q_0 + q_{n-1})^s - (q_0 + q_n)^s + q_0^s).
\end{aligned}$$

https://doi.org/10.1239/jap/1437658606 Published online by Cambridge University Press

Note that this relation is not true if at least one of the k_ℓ is 0. Indeed, if $k_\ell = 0$ for some $\ell = 1, \ldots, n-2$, we have $\mathbb{P}\{T_{n,n}(\boldsymbol{p}) \le k, N_1^{(k)} = k_1, \ldots, N_{n-2}^{(k)} = k_{n-2}\} = 0$. Summing over all the $k_1, \ldots, k_{n-2}$ such that $\sum_{i=1}^{n-2} k_i \le k$, we obtain

$$\mathbb{P}\{T_{n,n}(\boldsymbol{p}) \le k\} = \sum_{\boldsymbol{k} \in E_{n-2}} \frac{k!\, p_1^{k_1} \cdots p_{n-2}^{k_{n-2}} (1 - (p_1 + \cdots + p_{n-2}))^s}{k_1! \cdots k_{n-2}!\, s!} \times (1 - (q_0 + q_{n-1})^s - (q_0 + q_n)^s + q_0^s), \tag{11}$$

where E_{n-2} is defined by $E_{n-2} = \{\boldsymbol{k} = (k_1, \ldots, k_{n-2}) \in (\mathbb{N}^*)^{n-2} \mid k_1 + \cdots + k_{n-2} \le k\}$ and $\mathbb{N}^*$ is the set of positive integers. Note that for $n = 2$, we have

$$\mathbb{P}\{T_{2,2}(\boldsymbol{p}) \le k\} = 1 - (p_0 + p_1)^k - (p_0 + p_2)^k + p_0^k.$$

Recall that $p_0 = 1 - \sum_{i=1}^{n} p_i$. By definition of p'_{n-1} and p'_n, we have for every $\lambda \in (0, 1)$, $p'_{n-1} + p'_n = p_{n-1} + p_n$. It follows that, by the definition of $\boldsymbol{p}'$,

$$p'_0 = 1 - (p_1 + \cdots + p_{n-2} + p'_{n-1} + p'_n) = 1 - (p_1 + \cdots + p_{n-2} + p_{n-1} + p_n) = p_0.$$

Suppose that we have $p_{n-1} < p_n$. This implies, by the definition of p'_{n-1} and p'_n, that $p_{n-1} < p'_{n-1}$, $p'_n < p_n$, that is, $q_{n-1} < q'_{n-1}$, $q'_n < q_n$, where

$$q'_{n-1} = \frac{p'_{n-1}}{p'_0 + p'_{n-1} + p'_n} = \frac{p'_{n-1}}{p_0 + p_{n-1} + p_n},$$
$$q'_n = \frac{p'_n}{p'_0 + p'_{n-1} + p'_n} = \frac{p'_n}{p_0 + p_{n-1} + p_n}.$$

In the same way, we have

$$q'_0 = \frac{p'_0}{p'_0 + p'_{n-1} + p'_n} = \frac{p_0}{p_0 + p_{n-1} + p_n} = q_0.$$

Thus, $q_0 + q_{n-1} < q'_0 + q'_{n-1}$, $q'_0 + q'_n < q_0 + q_n$. The function $f(x) = x^s$ is convex on the interval $[0, 1]$ so, from Lemma 3, since $2q_0 + q_{n-1} + q_n = 2q'_0 + q'_{n-1} + q'_n$, we have

$$(q'_0 + q'_{n-1})^s + (q'_0 + q'_n)^s \le (q_0 + q_{n-1})^s + (q_0 + q_n)^s. \tag{12}$$

Similarly, if $p_n < p_{n-1}$, we have $p_n < p'_n$, $p'_{n-1} < p_{n-1}$, that is, $q_n < q'_n$, $q'_{n-1} < q_{n-1}$ and, thus, we also have (12) in this case. Substituting (12) into (11), we obtain, since $q'_0 = q_0$,

$$\begin{aligned}
&\mathbb{P}\{T_{n,n}(\boldsymbol{p}) \le k\} \\
&\quad \le \sum_{\boldsymbol{k} \in E_{n-2}} \frac{k!\, p_1^{k_1} \cdots p_{n-2}^{k_{n-2}} (1 - (p_1 + \cdots + p_{n-2}))^s}{k_1! \cdots k_{n-2}!\, s!} (1 - (q'_0 + q'_{n-1})^s - (q'_0 + q'_n)^s + q_0'^s) \\
&\quad = \mathbb{P}\{T_{n,n}(\boldsymbol{p}') \le k\},
\end{aligned}$$

which completes the proof.

https://doi.org/10.1239/jap/1437658606 Published online by Cambridge University Press

Theorem 3 can easily be extended to the case where the two entries p_{n-1} and p_n of $\boldsymbol{p}$, which are different from the entries p'_{n-1} and p'_n of $\boldsymbol{p}'$, are any $p_i, p_j \in \{p_1, \dots, p_n\}$, with $i \neq j$. This is due to the fact that the function $\mathbb{P}\{T_{n,n}(\boldsymbol{p}) \leq k\}$, as a function of $\boldsymbol{p}$, is symmetric.

In fact, we have shown in Theorem 3 that for fixed n and k, the function of $\boldsymbol{p}$, $\mathbb{P}\{T_{n,n}(\boldsymbol{p}) \leq k\}$, is a Schur-convex function, that is, a function that preserves the order of majorization; see [6].

Theorem 4. *For every $n \geq 1$ and $\boldsymbol{p} = (p_1, \dots, p_n) \in (0,1)^n$ with $\sum_{i=1}^n p_i \leq 1$, we have*

$$\mathbb{P}\{T_{n,n}(\boldsymbol{p}) > k\} \geq \mathbb{P}\{T_{n,n}(\boldsymbol{v}) > k\} \geq \mathbb{P}\{T_{n,n}(\boldsymbol{u}) > k\} \quad \textit{for every } k \geq 0,$$

where $\boldsymbol{u} = (1/n, \dots, 1/n)$, $\boldsymbol{v} = (v_1, \dots, v_n)$ with $v_i = (1 - p_0)/n$ and $p_0 = 1 - \sum_{i=1}^n p_i$.

Proof. To prove the first inequality, we apply Theorem 3 successively and at most $n - 1$ times as follows. We first choose two different entries of $\boldsymbol{p}$, say p_i and p_j such that $p_i < (1 - p_0)/n < p_j$ and then define p'_i and p'_j by

$$p'_i = \frac{1 - p_0}{n}, \qquad p'_j = p_i + p_j - \frac{1 - p_0}{n}.$$

From this we can write $p'_i = \lambda p_i + (1 - \lambda) p_j$ and $p'_j = (1 - \lambda) p_i + \lambda p_j$, with

$$\lambda = \frac{p_j - (1 - p_0)/n}{p_j - p_i}.$$

From Theorem 3, vector $\boldsymbol{p}'$, which is obtained by taking the other entries equal to those of $\boldsymbol{p}$, i.e. by taking $p'_\ell = p_\ell$ for $\ell \neq i, j$, is such that $\mathbb{P}\{T_{n,n}(\boldsymbol{p}) > k\} \geq \mathbb{P}\{T_{n,n}(\boldsymbol{p}') > k\}$. Note that at this point vector $\boldsymbol{p}'$ has at least one entry equal to $(1 - p_0)/n$, so repeating this procedure at most $n - 1$ times, we obtain vector $\boldsymbol{v}$.

To prove the second inequality we use (10). Introducing, for every $n \geq 1$, the set $F_n(\ell)$ defined by $F_n(\ell) = \{(k_1, \dots, k_n) \in (\mathbb{N}^*)^n \mid k_1 + \cdots + k_n = \ell\}$. For $k < n$, both terms are 0, so we suppose that $k \geq n$. We have

$$\begin{aligned}
\mathbb{P}\{T_{n,n}(\boldsymbol{v}) \leq k\} &= \mathbb{P}\{N_1^{(k)} > 0, \dots, N_n^{(k)} > 0\} \\
&= \sum_{k_0=0}^{k-n} \mathbb{P}\{N_0^{(k)} = k_0,\ N_1^{(k)} > 0, \dots, N_n^{(k)} > 0\} \\
&= \sum_{k_0=0}^{k-n} \sum_{(k_1,\dots,k_n) \in F_n(k-k_0)} \frac{k!}{k_0!\, k_1! \cdots k_n!} p_0^{k_0} \left(\frac{1 - p_0}{n}\right)^{k-k_0} \\
&= \sum_{k_0=0}^{k-n} \binom{k}{k_0} p_0^{k_0} (1 - p_0)^{k-k_0} \frac{1}{n^{k-k_0}} \sum_{(k_1,\dots,k_n) \in F_n(k-k_0)} \frac{(k - k_0)!}{k_1! \cdots k_n!}.
\end{aligned}$$

Setting $p_0 = 0$, we obtain

$$\mathbb{P}\{T_{n,n}(\boldsymbol{u}) \leq k\} = \frac{1}{n^k} \sum_{(k_1,\dots,k_n) \in F_n(k)} \frac{k!}{k_1! \cdots k_n!}.$$

https://doi.org/10.1239/jap/1437658606 Published online by Cambridge University Press

It follows that

$$\begin{aligned}\mathbb{P}\{T_{n,n}(\boldsymbol{v}) \le k\} &= \sum_{k_0=0}^{k-n} \binom{k}{k_0} p_0^{k_0}(1-p_0)^{k-k_0}\mathbb{P}\{T_{n,n}(\boldsymbol{u}) \le k-k_0\} \\ &\le \mathbb{P}\{T_{n,n}(\boldsymbol{u}) \le k\} \sum_{k_0=0}^{k-n} \binom{k}{k_0} p_0^{k_0}(1-p_0)^{k-k_0} \\ &\le \mathbb{P}\{T_{n,n}(\boldsymbol{u}) \le k\},\end{aligned}$$

which completes the proof.

To illustrate the steps used in the proof of this theorem, we provide the following example. Suppose that $n = 5$ and $\boldsymbol{p} = (\frac{1}{16}, \frac{1}{6}, \frac{1}{4}, \frac{1}{8}, \frac{7}{24})$. This implies that $p_0 = \frac{5}{48}$ and $(1-p_0)/n = \frac{43}{240}$. In the first step, taking $i = 4$ and $j = 5$, we obtain

$$\boldsymbol{p}^{[1]} = \left(\tfrac{1}{16}, \tfrac{1}{6}, \tfrac{1}{4}, \tfrac{43}{240}, \tfrac{19}{80}\right).$$

In the second step, taking $i = 2$ and $j = 5$, we obtain

$$\boldsymbol{p}^{[2]} = \left(\tfrac{1}{16}, \tfrac{43}{240}, \tfrac{1}{4}, \tfrac{43}{240}, \tfrac{9}{40}\right).$$

In the third step, taking $i = 1$ and $j = 3$, we obtain

$$\boldsymbol{p}^{[3]} = \left(\tfrac{43}{240}, \tfrac{43}{240}, \tfrac{2}{15}, \tfrac{43}{240}, \tfrac{9}{40}\right).$$

For the fourth and final step, taking $i = 5$ and $j = 3$, we obtain

$$\boldsymbol{p}^{[4]} = \left(\tfrac{43}{240}, \tfrac{43}{240}, \tfrac{43}{240}, \tfrac{43}{240}, \tfrac{43}{240}\right) = \tfrac{43}{48}\left(\tfrac{1}{5}, \tfrac{1}{5}, \tfrac{1}{5}, \tfrac{1}{5}, \tfrac{1}{5}\right).$$

6. A new conjecture

We propose a new conjecture stating that the complementary distribution function of $T_{c,n}$ is minimal when the distribution $\boldsymbol{p}$ is equal to the uniform distribution $\boldsymbol{u}$.

Conjecture 1. *For every $n \ge 1$, $c = 1, \ldots, n$ and $\boldsymbol{p} = (p_1, \ldots, p_n) \in (0,1)^n$ with $\sum_{i=1}^n p_i \le 1$, we have for all $k \ge 0$,*

$$\mathbb{P}\{T_{c,n}(\boldsymbol{p}) > k\} \ge \mathbb{P}\{T_{c,n}(\boldsymbol{v}) > k\} \ge \mathbb{P}\{T_{c,n}(\boldsymbol{u}) > k\},$$

where $\boldsymbol{u} = (1/n, \ldots, 1/n)$, $\boldsymbol{v} = (v_1, \ldots, v_n)$ with $v_i = (1-p_0)/n$ and $p_0 = 1 - \sum_{i=1}^n p_i$.

This new conjecture is motivated by the following facts:

- the result is true for the expectations; see Theorem 2,
- the result is true for $c = n$; see Theorem 4,
- the result is trivially true for $c = 1$ since

$$\mathbb{P}\{T_{1,n}(\boldsymbol{p}) > k\} = \mathbb{P}\{T_{1,n}(\boldsymbol{v}) > k\} = p_0^k \ge \mathbf{1}_{\{k=0\}} = \mathbb{P}\{T_{1,n}(\boldsymbol{u}) > k\},$$

- the result is true for $c = 2$; see Theorem 5 below.

https://doi.org/10.1239/jap/1437658606 Published online by Cambridge University Press

Theorem 5. *For every $n \geq 2$ and $\boldsymbol{p} = (p_1, \ldots, p_n) \in (0,1)^n$ with $\sum_{i=1}^n p_i \leq 1$, we have for every $k \geq 0$, $\mathbb{P}\{T_{2,n}(\boldsymbol{p}) > k\} \geq \mathbb{P}\{T_{2,n}(\boldsymbol{v}) > k\} \geq \mathbb{P}\{T_{2,n}(\boldsymbol{u}) > k\}$, where $\boldsymbol{u} = (1/n, \ldots, 1/n)$, $\boldsymbol{v} = (v_1, \ldots, v_n)$ with $v_i = (1 - p_0)/n$ and $p_0 = 1 - \sum_{i=1}^n p_i$.*

Proof. From (1), we have

$$\mathbb{P}\{T_{2,n}(\boldsymbol{p}) > k\} = -(n-1)p_0^k + \sum_{\ell=1}^{n} (p_0 + p_\ell)^k$$

and

$$\mathbb{P}\{T_{2,n}(\boldsymbol{v}) > k\} = -(n-1)p_0^k + n\left(p_0 + \frac{1-p_0}{n}\right)^k.$$

For every constant $a \geq 0$, the function $f(x) = (a+x)^k$ is convex on the interval $[0, \infty)$, so we have, taking $a = p_0$, by the Jensen inequality,

$$\left(p_0 + \frac{1-p_0}{n}\right)^k = \left(\frac{1}{n}\sum_{\ell=1}^{n}(p_0 + p_\ell)\right)^k \leq \frac{1}{n}\sum_{\ell=1}^{n}(p_0 + p_\ell)^k.$$

This implies that $\mathbb{P}\{T_{2,n}(\boldsymbol{p}) > k\} \geq \mathbb{P}\{T_{2,n}(\boldsymbol{v}) > k\}$.

To prove the second inequality, we define the function $F_{n,k}$ on the interval $[0, 1]$ by

$$F_{n,k}(x) = -(n-1)x^k + n\left(x + \frac{1-x}{n}\right)^k.$$

We then have $F_{n,k}(p_0) = \mathbb{P}\{T_{2,n}(\boldsymbol{v}) > k\}$ and $F_{n,k}(0) = \mathbb{P}\{T_{2,n}(\boldsymbol{u}) > k\}$. The derivative of function $F_{n,k}$ is

$$F'_{n,k}(x) = k(n-1)\left[\left(x + \frac{1-x}{n}\right)^{k-1} - x^{k-1}\right] \geq 0.$$

Function $F_{n,k}$ is, thus, an increasing function, which means that

$$\mathbb{P}\{T_{2,n}(\boldsymbol{v}) > k\} \geq \mathbb{P}\{T_{2,n}(\boldsymbol{u}) > k\},$$

which completes the proof.

7. Application to the detection of distributed deny of service attacks

A deny of service (DoS) attack tries to progressively take down an internet resource by flooding this resource with more requests than it is capable of handling. A distributed deny of service (DDoS) attack is a DoS attack triggered by thousands of machines that have been infected by malicious software, with as immediate consequence the total shut down of targeted web resources (e.g. e-commerce websites). A solution to detect and to mitigate DDoS attacks it to monitor network traffic at routers and to look for highly frequent signatures that might suggest ongoing attacks. A recent strategy followed by attackers is to hide their massive flow of requests over a multitude of routes, so that locally, these flows do not appear as frequent, while globally they represent a significant portion of the network traffic. The term 'iceberg' has been recently introduced to describe such an attack as only a very small part of the iceberg can be observed from each single router. The approach adopted to defend against such new attacks

https://doi.org/10.1239/jap/1437658606 Published online by Cambridge University Press

is to rely on multiple routers that locally monitor their network traffic, and upon detection of potential icebergs, inform a monitoring server that aggregates all the monitored information to accurately detect icebergs. Now to prevent the server from being overloaded by all the monitored information, routers continuously keep track of the c (among n) most recent high flows (modelled as items) prior to sending them to the server, and throw away all the items that appear with a small probability p_i, and such that the sum of these small probabilities is modelled by probability $1 - p_0$. Parameter c is dimensioned so that the frequency at which all the routers send their c last frequent items is low enough to enable the server to aggregate all of them and to trigger a DDoS alarm when needed. This amounts to computing the time needed to collect c distinct items among n frequent ones. Moreover, in Theorem 5 we have shown that the expectation of this time is minimal when the distribution of the frequent items is uniform.

References

[1] Boneh, A. and Hofri, M. (1997). The coupon-collector problem revisited—a survey of engineering problems and computational methods. *Commun. Statist. Stoch. Models* **13,** 39–66.

[2] Brown, M., Peköz, E. A. and Ross, S. M. (2008). Coupon collecting. *Prob. Eng. Inf. Sci.* **22,** 221–229.

[3] Doumas, A. V. and Papanicolaou, V. G. (2012). The coupon collector's problem revisited: asymptotics of the variance. *Adv. Appl. Prob.* **44,** 166–195.

[4] Doumas, A. V. and Papanicolaou, V. G. (2013). Asymptotics of the rising moments for the coupon collector's problem. *Electron. J. Prob.* **18,** 15pp.

[5] Flajolet, P., Gardy, D. and Thimonier, L. (1992). Birthday paradox, coupon collectors, caching algorithms and self-organizing search. *Discrete Appl. Math.* **39,** 207–229.

[6] Marshall, A. W., Olkin, I. and Arnold B. C. (2011). *Inequalities: Theory of Majorization and Its Applications*. 2nd. edn. Springer-Verlag, New York.

[7] Neal, P. (2008). The generalised coupon collector problem. *J. Appl. Prob.* **45,** 621–629.

[8] Rubin, H. and Zidek, J. (1965). A waiting time distribution arising from the coupon collector's problem. Tech. Rep. 107, Department of Statistics, Stanford University.

[9] Sericola, B. (2013). *Markov Chains: Theory, Algorithms and Applications*. Wiley-ISTE, London.