J. Appl. Probab. **60,** 723–736 (2023)
*doi:*10.1017/jpr.2022.94

THE NUMBER OF *K*-TONS IN THE COUPON COLLECTOR PROBLEM

JOHN C. SAUNDERS ,* *Middle Tennessee State University*

Abstract

Consider the coupon collector problem where each box of a brand of cereal contains a coupon and there are n different types of coupons. Suppose that the probability of a box containing a coupon of a specific type is $1/n$, and that we keep buying boxes until we collect at least m coupons of each type. For $k \geq m$ call a certain coupon a k-ton if we see it k times by the time we have seen m copies of all of the coupons. Here we determine the asymptotic distribution of the number of k-tons after we have collected m copies of each coupon for any k in a restricted range, given any fixed m. We also determine the asymptotic joint probability distribution over such values of k, and the total number of coupons collected.

Keywords: Probability distributions; combinatorial probability

2020 Mathematics Subject Classification: Primary 60E05
Secondary 60C05

1. Introduction

Consider the coupon collector problem where each box of a brand of cereal contains a coupon and there are n different types of coupons. Suppose that the probability of a box containing a coupon of a specific type is $1/n$, and that we keep buying boxes until we collect at least one coupon of each type. It is well known (see, for example, [9]) that the expected number of boxes we need to buy is nH_n, where H_n is the nth harmonic number:

$$H_n = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} = \log n + \gamma + \frac{1}{2n} + O\left(\frac{1}{n^2}\right),$$

where γ is the Euler–Mascheroni constant. The expected value of the total number of boxes collected has been extensively studied. For instance, it was proved in [9] that, if you continue to collect boxes until you have at least m coupons of each type, then the expected number of boxes collected is $n \log n + (m-1)n \log\log n + n \cdot C_m + o(n)$, where C_m is a constant depending on m.

Definition 1.1. The Gumbel distribution with parameters μ and β, which is denoted as Gumbel(μ, β), is defined as the probability distribution with cumulative density function $\exp[-e^{-(x-\mu)/\beta}]$ for $-\infty < x < \infty$.

The above result was improved in [5], which proved that, if $v_m(n)$ is the total number of boxes collected, then $(v_m(n)/n) - \log n - (m-1)\log\log n \sim \text{Gumbel}(-\log((m-1)!), 1)$ as

Received 13 August 2021; revision received 7 August 2022.
* Postal address: Middle Tennessee State University Department of Mathematical Sciences. Email: John.Saunders@mtsu.edu

$n \to \infty$. The probability distribution of the number of boxes that need to be collected to collect $a_n + 1$ coupons, where $0 \le a_n < n$, was studied in [1]. The expected number of boxes needing to be collected if the types of coupons have different probabilities of being in a box was examined in [2], and [3, 8] looked at the actual probability distribution of the number of boxes collected if the types of coupons have different probabilities. The case of collecting at least m coupons of each type has also been examined [4]. Further variations of the problem, where you have multiple collectors, were studied in [6, 7].

Here we examine the situation where we continue to collect boxes until we have at least m copies of each type of coupon for any fixed $m \in \mathbb{N}$.

Definition 1.2. For all $k \ge m$, call a certain coupon a k-ton if we see it k times by the time we have seen m copies of all of the coupons. Let the number of k-tons be denoted as S_k.

In the case of $m = 1$, [7] determined the expected number of 1-tons, which they called singletons, to be H_n. Singletons were also studied in [10], but in the case of unequal coupon probabilities and where the number of coupons collected in total is proportional to the number of types of coupons; many applications to their study, such as in databases, biological particles, and communication channels, were noted here. We extend this singleton result by determining the asymptotic joint probability distribution over values of k in a restricted range and the total number of coupons collected assuming, as in [7], equal coupon probabilities. We also determine the asymptotic distribution of the number of k-tons after we have collected m copies of each coupon for any such values of k, given any fixed $m \in \mathbb{N}$. While making an analogous extension to [10] is certainly interesting, we leave it for a future project.

Theorem 1.1. *Fix $m \in \mathbb{N}$, let $T_m(n)$ be the total number of coupons collected to see at least m copies of each coupon, and let $n \to \infty$. Let $k = o(\log n)$ if $m = 1$ and $k = o((\log n)/(\log\log n))$ if $m \ge 2$, with $k \ge m$ in either case. Then the joint probability distribution of*

$$\left(\frac{T_m(n) - n\log n - (m-1)n\log\log n}{n}, \frac{k!S_k}{(\log n)^{k-m+1}} \right)$$

converges to (X, e^{-X}), where $X \sim \mathrm{Gumbel}(-\log((m-1)!), 1)$.

A simple heuristic argument for Theorem 1.1 may be given as follows. First, from [5], we know that

$$\frac{T_m(n) - n\log n - (m-1)n\log\log n}{n} \sim \mathrm{Gumbel}(-\log((m-1)!), 1).$$

Later on, we also see, unsurprisingly, that the number of times a specific type of coupon is seen is essentially independent of whether or not all types of coupons have been collected at least m times. Also, unsurprisingly, it turns out that the number of times two different types of coupons are seen is asymptotically independent. Since the probability of a type of coupon being seen k times is

$$\binom{T_m(n)}{k}\left(\frac{1}{n}\right)^k\left(1-\frac{1}{n}\right)^{T_m(n)-k} \sim \frac{(n\log n)^k \mathrm{e}^{-\log n-(m-1)\log\log n-x}}{k!n^k} = \frac{(\log n)^{k-m+1}}{n\mathrm{e}^x k!},$$

where $x = (T_m(n) - n\log n - (m-1)n\log\log n)/n$, the expected number of k-tons is $\sim (\log n)^{k-m+1}/\mathrm{e}^x k!$, so that $k!S_k/(\log n)^{k-m+1} \sim \mathrm{e}^{-x}$.

https://doi.org/10.1017/jpr.2022.94 Published online by Cambridge University Press

While Theorem 1.1 provides us with the joint distribution between the total number of coupons collected and the number of k-tons for minimal k, we can also ask what happens for maximum values of k. The answer to this is provided in the next theorem.

Theorem 1.2. *Fix $m \in \mathbb{N}$ and let $T_m(n)$ be the total number of coupons collected to see at least m copies of each coupon.*

(i) Fix $d \in \mathbb{R}$. Pick an increasing sequence $(n_j)_j \subset \mathbb{N}$ and a sequence $(d_j)_j \subset \mathbb{R}$ such that $\lim_{j\to\infty} d_j = d$ and $k_j = \mathrm{e}\log n_j + \left((\mathrm{e}-1)(m-1) - \frac{1}{2}\right)\log\log n_j + d_j \in \mathbb{N}$ for all $j \in \mathbb{N}$. Then the joint probability distribution of

$$\left(\frac{T_m(n_j) - n_j \log n_j - (m-1)n_j \log\log n_j}{n_j}, S_{k_j}\right)$$

converges to

$$\left(X, \mathrm{Pois}\left(\frac{\mathrm{e}^{(\mathrm{e}-1)X-d}}{\sqrt{2\pi\mathrm{e}}}\right)\right),$$

where $X \sim \mathrm{Gumbel}(-\log((m-1)!), 1)$.

(ii) Let $g(n) = o(\log\log n)$ with $\lim_{n\to\infty} g(n) = \infty$, such that $k = \mathrm{e}\log n + \left((\mathrm{e}-1)(m-1) - \frac{1}{2}\right)\log\log n + g(n) \in \mathbb{N}$ for all $n \in \mathbb{N}$. Then S_k converges to 0 with probability 1.

(iii) Let $g(n) = o(\log\log n)$ with $\lim_{n\to\infty} g(n) = \infty$, such that $k = \mathrm{e}\log n + \left((\mathrm{e}-1)(m-1) - \frac{1}{2}\right)\log\log n - g(n) \in \mathbb{N}$ for all $n \in \mathbb{N}$. Then the value of $\mathbb{P}(S_k = 0)$ converges to 0.

A simple heuristic argument for Theorem 1.2 may be given as follows. The probability of a type of coupon being seen k_j times is

$$\binom{T_m(n_j)}{k_j}\left(\frac{1}{n_j}\right)^{k_j}\left(1-\frac{1}{n_j}\right)^{T_m(n_j)-k_j}.$$

Assuming that $x = (T_m(n_j) - n_j\log n_j - (m-1)n_j\log\log n_j)/n_j$ is constant as $j \to \infty$ and k_j is as defined in Theorem 1.2, we have that the above is asymptotic to

$$\begin{aligned}
&\frac{1}{\sqrt{2\pi k_j}}\left(\frac{\mathrm{e}T_m(n_j)}{n_j k_j}\right)^{k_j}\exp\left[-\frac{T_m(n_j)}{n_j}\right] \\
&\qquad\sim \frac{1}{\sqrt{2\pi\mathrm{e}\log n_j}}\exp\left[\frac{T_m(n_j)\mathrm{e} - k_j n_j - T_m(n_j)}{n_j}\right] \\
&\qquad= \frac{1}{\sqrt{2\pi\mathrm{e}\log n_j}}\exp\left[-\log n_j + \frac{\log\log n_j}{2} + (\mathrm{e}-1)x - d_j\right] \\
&\qquad= \frac{1}{\sqrt{2\pi\mathrm{e}n_j}}\mathrm{e}^{(\mathrm{e}-1)x-d_j}.
\end{aligned}$$

Therefore, by similar reasoning to the heuristic for Theorem 1.1, we can see that the probability that there are y k-tons is asymptotic to

$$\binom{n}{y}\left(\frac{\lambda}{n}\right)^y\left(1-\frac{\lambda}{n}\right)^{n-y} \sim \frac{\lambda^y\mathrm{e}^{-\lambda}}{y!},$$

https://doi.org/10.1017/jpr.2022.94 Published online by Cambridge University Press

where $\lambda = e^{(e-1)x-d}/(\sqrt{2\pi e})$, which gives the Poisson distribution in Theorem 1.2(i). Parts (ii) and (iii) can be heuristically argued by letting d_j approach ∞ or $-\infty$.

From Theorem 1.1 we can derive the following corollaries.

Corollary 1.1. *Fix $m \in \mathbb{N}$, let $n \to \infty$, and let $k = o(\log n)$ if $m = 1$ and $k = o((\log n)/(\log\log n))$ if $m \geq 2$, with $k \geq m$ in either case. Suppose we keep collecting coupons until we see at least m copies of each coupon. Then*

$$\frac{k!S_k}{(\log n)^{k-m+1}} \xrightarrow{D} \operatorname{Exp}\left(\frac{1}{(m-1)!}\right).$$

Corollary 1.1 gives the asymptotic probability distribution of the number of k-tons. If we fix k, we can also determine the joint asymptotic distribution between the number of m-tons, $m+1$-tons, ..., and k-tons. We can see that asymptotically all of these variables are linearly dependent.

Corollary 1.2. *Fix $k, m \in \mathbb{N}$ with $k \geq m$, and let $n \to \infty$. Suppose we keep collecting coupons until we see at least m copies of each coupon. Then the joint probability distribution*

$$\left(\frac{m!S_m}{(\log n)}, \frac{(m+1)!S_{m+1}}{(\log n)^2}, \ldots, \frac{k!S_k}{(\log n)^{k-m+1}}\right)$$

converges to $X \cdot (1, 1, \ldots, 1)$, where

$$X \sim \operatorname{Exp}\left(\frac{1}{(m-1)!}\right).$$

2. Proofs

To prove Theorem 1.1 we require the following notation and lemma.

Note 2.1. Suppose we collect x coupons in total, regardless of whether we have collected all n types of coupons or not. Let $S_{i,x}$ denote the number of types of coupons we collected exactly i times.

Note 2.2. Throughout this paper, let $N(n, f(n)) := n\log n + (m-1)n\log\log n + f(n)$.

Lemma 2.1. *Let $f: \mathbb{N} \to \mathbb{R}$ be such that $-\frac{1}{2}n\log\log n \leq f(n) \leq \frac{1}{2}n\log\log n$, and $N(n, f(n)) \in \mathbb{N}$ for all $n \in \mathbb{N}$. Suppose we collect N(n,f(n)) coupons in total, regardless of whether we collect m copies of all n different kinds of coupons or not. Let $k = o(\log n)$ if $m = 1$ and $k = o(\log n/\log\log n)$ if $m \geq 2$. If $m \leq k$, then as $n \to \infty$*

$$\frac{k!S_{k,N(n,f(n))}}{(\log n)^{k-m+1}}$$

is asymptotic to $e^{-f(n)/n}$ with probability 1, and

$$\operatorname{Var}\left(\frac{k!S_{k,N(n,f(n))}}{(\log n)^{k-m+1}}\right) = O\left(\frac{1}{\sqrt{\log n}}\right).$$

https://doi.org/10.1017/jpr.2022.94 Published online by Cambridge University Press

Proof. First, assume that $m \le k$. We can see that, as $n \to \infty$,

$$\begin{aligned}\mathbb{E}(S_{k,N(n,f(n))}) &= n \cdot \mathbb{P}(c_0 \text{ occurs exactly } k \text{ times in the collection of } N(n,f(n)) \text{ coupons})\\ &= \binom{N(n,f(n))}{k}\left(1-\frac{1}{n}\right)^{N(n,f(n))-k}\frac{1}{n^{k-1}}\\ &\sim \frac{N(n,f(n))^k}{k!n^{k-1}}\exp\left[-\log n-(m-1)\log\log n-\frac{f(n)}{n}\right]\\ &\sim \frac{(\log n)^{k-m+1}}{k!\mathrm{e}^{f(n)/n}}.\end{aligned}$$

Thus, we can deduce that

$$\mathbb{E}\left(\frac{k!S_{k,N(n,f(n))}}{(\log n)^{k-m+1}}\right) = \mathrm{e}^{-f(n)/n}(1+o(1))$$

as $n \to \infty$. Also,

$$\begin{aligned}&\mathbb{E}(S^2_{k,N(n,f(n))})\\ &\quad= n(n-1)\cdot\mathbb{P}(c_0, c_1 \text{ both occur exactly } k \text{ times in the collection of } N(n,f(n)) \text{ coupons})\\ &\qquad+ n\cdot\mathbb{P}(c_0 \text{ occurs exactly } k \text{ times in the collection of } N(n,f(n)) \text{ coupons})\\ &\quad= n(n-1)\binom{N(n,f(n))}{k}\binom{N(n,f(n))-k}{k}\left(1-\frac{2}{n}\right)^{N(n,f(n))-2k}\frac{1}{n^{2k}}\\ &\qquad+\binom{N(n,f(n))}{k}\left(1-\frac{1}{n}\right)^{N(n,f(n))-k}\frac{1}{n^{k-1}}\\ &\quad\le (\mathbb{E}(S_{k,N(n,f(n))}))^2\left(1-\frac{1}{n}\right)^{-2k}+\mathbb{E}(S_{k,N(n,f(n))}).\end{aligned} \tag{2.1}$$

Notice that

$$\frac{k!\mathrm{e}^{-f(n)/n}}{(\log n)^{k-m+1}} \le \frac{k!}{(\log n)^{k-m+\frac{1}{2}}}. \tag{2.2}$$

If k is bounded as $n \to \infty$, then the right-hand side of (2.2) $O(1/\sqrt{\log n})$. On the other hand, for k sufficiently large we have $k! < k^{k-m}$, so even if values of k tend toward ∞, the right-hand side of (2.2) is $O(1/\sqrt{\log n})$. Thus, from (2.1), we obtain

$$\begin{aligned}&\mathbb{E}\left(\left(\frac{k!S_{k,N(n,f(n))}}{(\log n)^{k-m+1}}\right)^2\right)-\left(\mathbb{E}\left(\frac{k!S_{k,N(n,f(n))}}{(\log n)^{k-m+1}}\right)\right)^2\\ &\quad\le\left(\mathbb{E}\left(\frac{k!S_{k,N(n,f(n))}}{(\log n)^{k-m+1}}\right)\right)^2\cdot O\left(\frac{k}{n}\right)+\frac{(k!)^2\mathbb{E}(S_{k,N(n,f(n))})}{(\log n)^{2k-2m+2}}\\ &\quad\le O\left(\frac{k\sqrt{\log n}}{n}\right)+O\left(\frac{1}{\sqrt{\log n}}\right)=O\left(\frac{1}{\sqrt{\log n}}\right)\end{aligned}$$

https://doi.org/10.1017/jpr.2022.94 Published online by Cambridge University Press

as $n \to \infty$. Thus,

$$\operatorname{Var}\left(\frac{k!S_{k,N(n,f(n))}}{(\log n)^{k-m+1}}\right) = O\left(\frac{1}{\sqrt{\log n}}\right)$$

as $n \to \infty$, so that we obtain our result for $m \le k$. □

Proof of Theorem 1.1. $X = (T_m(n) - n\log n - (m-1)n\log\log n)/n$ was proved to be asymptotically Gumbel distributed in [5]. To prove Theorem 1.1, it therefore suffices to show that

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{T_m(n) - n\log n - (m-1)n\log\log n}{n} < x,\ \frac{k!S_k}{(\log n)^{k-m+1}} < \mathrm{e}^{-y}\right) = 0 \tag{2.3}$$

for any fixed $x < y$, and that

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{T_m(n) - n\log n - (m-1)n\log\log n}{n} > x,\ \frac{k!S_k}{(\log n)^{k-m+1}} > \mathrm{e}^{-y}\right) = 0 \tag{2.4}$$

for any fixed $y < x$. To prove (2.3), we first calculate an upper bound for

$$\mathbb{P}\left(T_m(n) \le N(n, nx),\ \frac{k!S_k}{(\log n)^{k-m+1}} < \mathrm{e}^{-y}\right)$$

for all sufficiently large $n \in \mathbb{N}$. Let $M(n) = n\log n + \left(m - \frac{3}{2}\right)n\log\log n$. We have

$$\begin{aligned}
&\mathbb{P}\left(T_m(n) \le N(n, nx),\ \frac{k!S_k}{(\log n)^{k-m+1}} < \mathrm{e}^{-y}\right)\\
&\quad= \mathbb{P}\left(T_m(n) < M(n),\ \frac{k!S_k}{(\log n)^{k-m+1}} < \mathrm{e}^{-y}\right)\\
&\qquad+ \mathbb{P}\left(M(n) \le T_m(n) \le N(n, nx),\ \frac{k!S_k}{(\log n)^{k-m+1}} < \mathrm{e}^{-y}\right)\\
&\quad\le \mathbb{P}(T_m(n) < M(n)) + \mathbb{P}\left(M(n) \le T_m(n) \le N(n, nx),\ \frac{k!S_k}{(\log n)^{k-m+1}} < \mathrm{e}^{-y}\right).
\end{aligned}$$

We have

$$\begin{aligned}
&\mathbb{P}\left(M(n) \le T_m(n) \le N(n, nx),\ \frac{k!S_k}{(\log n)^{k-m+1}} < \mathrm{e}^{-y}\right)\\
&\qquad= \sum_{M=\lceil M(n)\rceil}^{\lfloor N(n,nx)\rfloor} \mathbb{P}\left(T_m(n) = M,\ \frac{k!S_k}{(\log n)^{k-m+1}} < \mathrm{e}^{-y}\right).
\end{aligned} \tag{2.5}$$

Clearly,

$$\begin{aligned}
\mathbb{P}\left(T_m(n) = M,\ \frac{k!S_k}{(\log n)^{k-m+1}} < \mathrm{e}^{-y}\right) &\le \mathbb{P}\left(S_{m-1,M-1} = 1, S_{m-1,M} = 0,\ \frac{k!S_{k,M}}{(\log n)^{k-m+1}} < \mathrm{e}^{-y}\right)\\
&= \frac{1}{n}\mathbb{P}\left(S_{m-1,M-1} = 1,\ \frac{k!S_{k,M}}{(\log n)^{k-m+1}} < \mathrm{e}^{-y}\right)\\
&\le \frac{1}{n}\mathbb{P}\left(\frac{k!S_{k,M}}{(\log n)^{k-m+1}} < \mathrm{e}^{-y}\right).
\end{aligned} \tag{2.6}$$

https://doi.org/10.1017/jpr.2022.94 Published online by Cambridge University Press

Combining (2.5) and (2.6), we therefore have

$$\mathbb{P}\left(M(n) \le T_m(n) \le N(n, nx), \frac{k!S_{k,M}}{(\log n)^{k-m+1}} < \mathrm{e}^{-y}\right)$$
$$\le \sum_{M=\lceil M(n)\rceil}^{\lfloor N(n,nx)\rfloor} \frac{1}{n}\mathbb{P}\left(\frac{k!S_{k,M}}{(\log n)^{k-m+1}} < \mathrm{e}^{-y}\right). \quad (2.7)$$

From Chebyshev's inequality and Lemma 2.1 we have, for sufficiently large n,

$$\mathbb{P}\left(\frac{k!S_{k,M}}{(\log n)^{k-m+1}} < \mathrm{e}^{-y}\right) \le \frac{C}{\sqrt{\log n}}$$

for some constant $C > 0$, depending only on x and y. From (2.7), we can thus deduce that

$$\mathbb{P}\left(M(n) \le T_m(n) \le N(n, nx), \frac{k!S_{k,M}}{(\log n)^{k-m+1}} < \mathrm{e}^{-y}\right)$$
$$\le \sum_{M=\lceil M(n)\rceil}^{\lfloor N(n,nx)\rfloor} \frac{C}{n\sqrt{\log n}}$$
$$\le \frac{C\left(nx + \frac{1}{2}n\log\log n\right)}{n\sqrt{\log n}} = \frac{C(2x + \log\log n)}{2\sqrt{\log n}} \quad (2.8)$$

for all sufficiently large n.

Now we let $n \to \infty$. From the result quoted at the start of the proof, we can deduce that $\lim_{n\to\infty} \mathbb{P}(T < M(n)) = 0$. From (2.8) we can also see that

$$\lim_{n\to\infty} \mathbb{P}\left(M(n) \le T_m(n) \le N(n, nx), \frac{k!S_{k,M}}{(\log n)^{k-m+1}} < \mathrm{e}^{-y}\right) = 0.$$

These two limits give us (2.3).

We can prove (2.4) similarly, replacing $M(n) = n\log n + \left(m - \frac{3}{2}\right)n\log\log n$ with $M'(n) = n\log n + \left(m - \frac{1}{2}\right)n\log\log n$ and reversing the appropriate inequalities. □

Proof of Corollaries 1.1 *and* 1.2. Corollaries 1.1 and 1.2 follow from Theorem 1.1 by showing that, if $X \sim \text{Gumbel}(-\log((m-1)!), 1)$, then

$$\mathrm{e}^{-X} \sim \text{Exp}\left(\frac{1}{(m-1)!}\right).$$

Indeed, we have, for any $r > 0$,

$$\mathbb{P}(\mathrm{e}^{-X} \le r) = \mathbb{P}(X \ge -\log r) = 1 - \mathbb{P}(X \le -\log r)$$
$$= 1 - \exp\left[-\mathrm{e}^{-(-\log r + \log(m-1)!)}\right]$$
$$= 1 - \mathrm{e}^{-r/(m-1)!}.$$

Hence, the result follows. □

To prove Theorem 1.2, we require the following lemma.

https://doi.org/10.1017/jpr.2022.94 Published online by Cambridge University Press

Lemma 2.2. *Fix $m \in \mathbb{N}$. Let $f\colon \mathbb{N} \to \mathbb{R}$ be such that there exists $x \in \mathbb{R}$ such that $\lim_{n\to\infty} f(n)/n = x$ and $N(n, f(n)) \in \mathbb{N}$ for all $n \in \mathbb{N}$. Suppose we collect $N(n, f(n))$ coupons in total regardless of whether we collect m copies of all n different kinds of coupons or not. Pick an increasing sequence $(n_j)_j \subset \mathbb{N}$ and a sequence $(d_j)_j \subset \mathbb{R}$ such that $\lim_{j\to\infty} d_j = d$ for some $d \in \mathbb{R}$ and $k = \mathrm{e}\log n_j + \big((\mathrm{e}-1)(m-1) - \frac{1}{2}\big)\log\log n_j + d_j \in \mathbb{N}$ for all $j \in \mathbb{N}$. Then, as $j \to \infty$,*

$$\begin{aligned}\mathbb{P}(S_{m-1,N(n_j,f(n_j))} = r_1,\, S_{k,N(n_j,f(n_j))} = r_2) = \frac{1}{r_1!r_2!}\left(\frac{\mathrm{e}^{-f(n_j)/n_j}}{(m-1)!}\right)^{r_1} \exp\left[\frac{-\mathrm{e}^{-f(n_j)/n_j}}{(m-1)!}\right]\\ \times \left(\frac{\exp[((\mathrm{e}-1)f(n_j)/n_j) - d]}{\sqrt{2\pi\mathrm{e}}}\right)^{r_2}\\ \times \exp\left[\frac{-\exp[((\mathrm{e}-1)f(n_j)/n_j) - d]}{\sqrt{2\pi\mathrm{e}}}\right]\\ \times (1 + o(1)),\end{aligned}$$

where the implied constant in the error term only depends upon r_1 and r_2.

Proof. Take some $j \in \mathbb{N}$. Let the probability that r_1 prescribed types of coupons are $(m-1)$-tons and r_2 prescribed types of coupons are k-tons be denoted as $W_{r_1,r_2}(n)$. Then, as $j \to \infty$,

$$\begin{aligned}&\binom{n_j}{r_1, r_2, n_j - r_1 - r_2} W_{r_1,r_2}(n_j)\\ &\quad= \binom{n_j}{r_1, r_2, n - r_1 - r_2} \frac{N(n_j, f(n_j))!}{(m-1)!^{r_1}k!^{r_2}(N(n_j, f(n_j)) - (m-1)r_1 - kr_2)!n_j^{(m-1)r_1+kr_2}}\\ &\qquad\times \left(1 - \frac{r_1 + r_2}{n_j}\right)^{N(n_j,f(n_j))-(m-1)r_1-kr_2}\\ &\quad= \binom{n_j}{r_1, r_2, n_j - r_1 - r_2} \frac{N(n_j, f(n_j))!\mathrm{e}^{-(r_1+r_2)N(n_j,f(n_j))/n_j}}{(m-1)!^{r_1}k!^{r_2}(N(n_j, f(n_j)) - (m-1)r_1 - kr_2)!n_j^{(m-1)r_1+kr_2}}\\ &\qquad\times \left(1 + O\left(\frac{\log n_j}{n_j}\right)\right)\\ &\quad= \frac{N(n_j, f(n_j))^{(m-1)r_1+kr_2}\mathrm{e}^{-(r_1+r_2)f(n_j)/n_j}}{r_1!r_2!(m-1)!^{r_1}k!^{r_2}(\log n_j)^{(r_1+r_2)(m-1)}n_j^{(m-1)r_1+kr_2}}\left(1 + O\left(\frac{\log n_j}{n_j}\right)\right)\\ &\quad= \frac{1}{r_1!r_2!}\left(\frac{\mathrm{e}^{-f(n_j)/n_j}}{(m-1)!}\right)^{r_1}\left(\frac{\mathrm{e}^{((\mathrm{e}-1)f(n_j)/n_j)-d_j}}{\sqrt{2\pi\mathrm{e}}}\right)^{r_2}\left(1 + O\left(\frac{(\log\log n_j)^2}{\log n_j}\right)\right),\end{aligned}$$

where the implied constant in the error term only depends on r_1 and r_2. By the inclusion–exclusion formula, we have

https://doi.org/10.1017/jpr.2022.94 Published online by Cambridge University Press

$$\mathbb{P}(S_{m-1,N(n_j,f(n_j))} = r_1, S_{k,N(n_j,f(n_j))} = r_2)$$
$$= \sum_{0\le j_1+j_2\le n_j-r_1-r_2} \left[(-1)^{j_1+j_2} \binom{n_j}{r_1+j_1, r_2+j_2, n_j-r_1-r_2-j_1-j_2} \binom{r_1+j_1}{r_1} \binom{r_2+j_2}{r_2} \right.$$
$$\left. \times W_{r_1+j_1,r_2+j_2}(n_j) \right]. \tag{2.9}$$

For $t \le n_j - r_1 - r_2$ even:

$$\mathbb{P}(S_{m-1,N(n_j,f(n_j))} = r_1, S_{k,N(n_j,f(n_j))} = r_2)$$
$$\le \sum_{0\le j_1+j_2\le t} \left[(-1)^{j_1+j_2} \binom{n_j}{r_1+j_1, r_2+j_2, n_j-r_1-r_2-j_1-j_2} \binom{r_1+j_1}{r_1} \binom{r_2+j_2}{r_2} \right.$$
$$\left. \times W_{r_1+j_1,r_2+j_2}(n_j) \right]. \tag{2.10}$$

For $t \le n_j - r_1 - r_2$ odd:

$$\mathbb{P}(S_{m-1,N(n_j,f(n_j))} = r_1, S_{k,N(n_j,f(n_j))} = r_2)$$
$$\ge \sum_{0\le j_1+j_2\le t} \left[(-1)^{j_1+j_2} \binom{n_j}{r_1+j_1, r_2+j_2, n_j-r_1-r_2-j_1-j_2} \binom{r_1+j_1}{r_1} \binom{r_2+j_2}{r_2} \right.$$
$$\left. \times W_{r_1+j_1,r_2+j_2}(n_j) \right]. \tag{2.11}$$

Combining (2.9)–(2.11), we have

$$\mathbb{P}(S_{m-1,N(n_j,f(n_j))} = r_1, S_{k,N(n_j,f(n_j))} = r_2)$$
$$\sim \frac{1}{r_1!r_2!} \sum_{j_1,j_2=0}^{\infty} \frac{(-1)^{j_1+j_2}}{j_1!j_2!} \left(\frac{\mathrm{e}^{-x}}{(m-1)!} \right)^{r_1+j_1} \left(\frac{\mathrm{e}^{((\mathrm{e}-1)x)-d}}{\sqrt{2\pi\mathrm{e}}} \right)^{r_2+j_2}$$
$$= \frac{1}{r_1!r_2!} \left(\frac{\mathrm{e}^{-x}}{(m-1)!} \right)^{r_1} \exp\left[\frac{-\mathrm{e}^{-x}}{(m-1)!} \right] \left(\frac{\mathrm{e}^{((e-1)x)-d}}{\sqrt{2\pi\mathrm{e}}} \right)^{r_2} \exp\left[\frac{-\mathrm{e}^{((e-1)x)-d}}{\sqrt{2\pi\mathrm{e}}} \right],$$

as $j \to \infty$. □

Proof of Theorem 1.2. We first prove Theorem 1.2(i). Let $d \in \mathbb{R}$ and pick increasing sequences $(n_j)_j$ and $(d_j)_j$ such that $\lim_{j\to\infty} d_j = d$ and $k_j := \mathrm{e}\log n_j + \big((\mathrm{e}-1)(m-1) - \frac{1}{2}\big)\log\log n_j + d_j \in \mathbb{N}$ for all $j \in \mathbb{N}$. Let $x < y$ and $r \in \mathbb{N}$ be constants. Consider the probability $\mathbb{P}(N(n_j, n_jx) \le T_m(n) \le N(n_j, n_jy), S_k = r)$. We have

$$\mathbb{P}(N(n_j, n_jx) \le T_m(n) \le N(n_j, n_jy), S_k = r) = \sum_{M=\lceil N(n_j,n_jx)\rceil}^{\lfloor N(n_j,n_jy)\rfloor} \mathbb{P}(T = M, S_{k,M-1} = r).$$

https://doi.org/10.1017/jpr.2022.94 Published online by Cambridge University Press

Clearly,

$$\begin{aligned}\mathbb{P}(T_m(n) = M,\, S_{k,M-1} = r) &\le \mathbb{P}(S_{m-1,M-1} = 1,\, S_{m-1,M} = 0,\, S_{k,M-1} = r)\\ &= \mathbb{P}(S_{m-1,M-1} = 1,\, S_{k,M-1} = r)/n. \end{aligned} \tag{2.12}$$

Also, for $M \ge N(n_j, n_j x)$, if we have $S_{i,M} > 0$ for some $0 \le i \le m-2$, then we have $S_{l,\lceil N(n_j,n_j x)\rceil} > 0$ for some $0 \le l \le i$. Thus, we can also deduce that

$$\begin{aligned}&\mathbb{P}(N(n_j, n_j x) \le T_m(n) \le N(n_j, n_j y),\, S_k = r)\\ &\quad\ge \left(\sum_{M=\lceil N(n_j,n_j x)\rceil}^{\lfloor N(n_j,n_j y)\rfloor} \frac{\mathbb{P}(S_{m-1,M-1} = 1,\, S_{k,M-1} = r)}{n_j}\right)\\ &\qquad - \mathbb{P}(\text{there exists } 0 \le l \le m-2 \text{ such that } S_{l,\lceil N(n_j,n_j x)\rceil} > 0). \end{aligned} \tag{2.13}$$

For any $0 \le l \le m-2$, notice that, as $j \to \infty$,

$$\begin{aligned}&\mathbb{E}(S_{l,\lceil N(n_j,n_j x)\rceil}) = n \cdot \mathbb{P}(c_0 \text{ occurs exactly } l \text{ times in the collection of } \lceil N(n_j, n_j x)\rceil \text{ coupons})\\ &\quad= \binom{\lceil N(n_j, n_j x)\rceil}{l}\left(1 - \frac{1}{n_j}\right)^{\lceil N(n_j,n_j x)\rceil - l} \frac{1}{n_j^{l-1}}\\ &\quad\sim \frac{N(n_j, n_j x)^l}{l! n_j^{l-1}} \mathrm{e}^{-\log n_j - (m-1)\log\log n_j - x} \sim \frac{(\log n_j)^{l-m+1}\mathrm{e}^{-x}}{l!}. \end{aligned}$$

Since $l \le m-2$, we therefore have $\lim_{j\to\infty} \mathbb{E}(S_{l,\lceil N(n_j,n_j x)\rceil}) = 0$, and so

$$\lim_{j\to\infty} \mathbb{P}(\text{there exists } 0 \le l \le m-2 \text{ such that } S_{l,\lceil N(n_j,n_j x)\rceil} > 0) = 0. \tag{2.14}$$

For every $N(n_j, n_j x) \le M \le N(n_j, n_j y)$, let $c_{M,j} := (M - n_j \log n_j - (m-1)n_j \log\log n_j)/n_j$. By Lemma 2.2, we have

$$\begin{aligned}&\sum_{M=\lceil N(n_j,n_j x)\rceil}^{\lfloor N(n_j,n_j y)\rfloor} \frac{\mathbb{P}(S_{m-1,M-1} = 1,\, S_{k,M-1} = r)}{n}\\ &= \sum_{M=\lceil N(n_j,n_j x)\rceil}^{\lfloor N(n_j,n_j y)\rfloor} \frac{1}{r! n}\left(\frac{\mathrm{e}^{-c_{M,j}}}{(m-1)!}\right)\exp\left[\frac{-\mathrm{e}^{-c_{M,j}}}{(m-1)!}\right]\left(\frac{\mathrm{e}^{(\mathrm{e}-1)c_{M,j}-d}}{\sqrt{2\pi\mathrm{e}}}\right)^r \exp\left[\frac{-\mathrm{e}^{(\mathrm{e}-1)c_{M,j}-d}}{\sqrt{2\pi\mathrm{e}}}\right]\\ &\quad\times(1 + o(1))\\ &< \left(\frac{\mathrm{e}^{(\mathrm{e}-1)y-d}}{\sqrt{2\pi\mathrm{e}}}\right)^r \exp\left[\frac{-\mathrm{e}^{(\mathrm{e}-1)x-d}}{\sqrt{2\pi\mathrm{e}}}\right]\\ &\quad\times \sum_{M=\lceil N(n_j,n_j x)\rceil}^{\lfloor N(n_j,n_j y)\rfloor} \frac{1}{r! n}\left(\frac{\mathrm{e}^{-c_{M,j}}}{(m-1)!}\right)\exp\left[\frac{-\mathrm{e}^{-c_{M,j}}}{(m-1)!}\right](1 + o(1)). \end{aligned} \tag{2.15}$$

https://doi.org/10.1017/jpr.2022.94 Published online by Cambridge University Press

Note that

$$\lim_{j\to\infty} \sum_{M=\lceil N(n_j,n_jx)\rceil}^{\lfloor N(n_j,n_jy)\rfloor} \frac{1}{n}\left(\frac{\mathrm{e}^{-c_{M,j}}}{(m-1)!}\right)\exp\left[\frac{-\mathrm{e}^{-c_{M,j}}}{(m-1)!}\right]$$

$$= \int_x^y \frac{\mathrm{e}^{-t}\exp\left[\frac{-\mathrm{e}^{-t}}{(m-1)!}\right]}{(m-1)!}\,\mathrm{d}\mathrm{d}t$$

$$= \exp\left[\frac{-\mathrm{e}^{-y}}{(m-1)!}\right] - \exp\left[\frac{-\mathrm{e}^{-x}}{(m-1)!}\right]$$

$$= \lim_{j\to\infty} \mathbb{P}(N(n_j, x) \le T_m(n_j) \le N(n_j, y)). \tag{2.16}$$

Combining (2.12)–(2.16), we obtain

$$\limsup_{j\to\infty} \mathbb{P}(N(n_j, n_jx) \le T_m(n_j) \le N(n_j, n_jy), S_k = r)$$

$$\le \frac{1}{r!}\left(\frac{\mathrm{e}^{(\mathrm{e}-1)y-d}}{\sqrt{2\pi\mathrm{e}}}\right)^r \exp\left[\frac{-\mathrm{e}^{(\mathrm{e}-1)x-d}}{\sqrt{2\pi\mathrm{e}}}\right] \lim_{j\to\infty} \mathbb{P}(N(n_j, n_jx) \le T_m(n_j) \le N(n_j, n_jy)).$$

By similar reasoning, we can also obtain

$$\liminf_{j\to\infty} \mathbb{P}(N(n_j, n_jx) \le T_m(n_j) \le N(n_j, n_jy), S_k = r)$$

$$\ge \frac{1}{r!}\left(\frac{\mathrm{e}^{(\mathrm{e}-1)x-d}}{\sqrt{2\pi\mathrm{e}}}\right)^r \exp\left[\frac{-\mathrm{e}^{(\mathrm{e}-1)y-d}}{\sqrt{2\pi\mathrm{e}}}\right] \lim_{j\to\infty} \mathbb{P}(N(n_j, n_jx) \le T_m(n_j) \le N(n_j, n_jy)).$$

Since the choices of $x < y$ were arbitrary, we can see that the joint limiting distribution holds. □

For Theorem 1.2(ii), again let $x < y$ be constants. We will show that

$$\lim_{n\to\infty} \mathbb{P}(N(n, nx) \le T_m(n_j) \le N(n, ny), S_k \ge 1) = 0,$$

which implies the desired result since $x < y$ are arbitrary. We have

$$\mathbb{P}(N(n, nx) \le T_m(n) \le N(n, ny), S_k \ge 1)$$

$$\le \sum_{M=\lceil N(n,nx)\rceil}^{\lfloor N(n,ny)\rfloor} \frac{\mathbb{P}(S_{m-1,M-1} = 1, S_{k,M-1} \ge 1)}{n}$$

$$\le \sum_{M=\lceil N(n,nx)\rceil}^{\lfloor N(n,ny)\rfloor} \frac{\mathbb{P}(S_{k,M-1} \ge 1)}{n}$$

$$\le \sum_{M=\lceil N(n,nx)\rceil}^{\lfloor N(n,ny)\rfloor} \frac{n\cdot\mathbb{P}(c_0 \text{ occurs exactly } k \text{ times in the collection of } M-1 \text{ coupons})}{n}$$

https://doi.org/10.1017/jpr.2022.94 Published online by Cambridge University Press

$$
\begin{aligned}
&= \sum_{M=\lceil N(n,nx)\rceil}^{\lfloor N(n,ny)\rfloor} \mathbb{P}(c_0 \text{ occurs exactly } k \text{ times in the collection of } M \text{ coupons}) \\
&= \sum_{M=\lceil N(n,nx)\rceil}^{\lfloor N(n,ny)\rfloor} \binom{M}{k}\left(1-\frac{1}{n}\right)^{M-k}\frac{1}{n^k} \\
&< \frac{1}{k!}\sum_{M=\lceil N(n,nx)\rceil}^{\lfloor N(n,ny)\rfloor}\left(1-\frac{1}{n}\right)^{M-k}\frac{M^k}{n^k} \\
&< \frac{n(y-x)}{k!}\left(1-\frac{1}{n}\right)^{N(n,nx)-k}(\log n+(m-1)\log\log n+y)^k.
\end{aligned}
$$

As $n\to\infty$, we have

$$
\begin{aligned}
&\frac{n(y-x)}{k!}\left(1-\frac{1}{n}\right)^{N(n,nx)-k}(\log n+(m-1)\log\log n+y)^k \\
&\quad\sim \frac{(y-x)(\log n+(m-1)\log\log n+y)^k}{k!(\log n)^{m-1}\mathrm{e}^x} \\
&\quad\sim \frac{y-x}{\sqrt{2\pi\mathrm{e}}(\log n)^{m-\frac{1}{2}}\mathrm{e}^x}\left(\frac{\mathrm{e}\log n+\mathrm{e}(m-1)\log\log n+\mathrm{e}y}{k}\right)^k \\
&\quad\sim \frac{(y-x)\mathrm{e}^{\left(m-\frac{1}{2}\right)\log\log n+\mathrm{e}y-g(n)}}{\sqrt{2\pi\mathrm{e}}(\log n)^{m-\frac{1}{2}}\mathrm{e}^x} \\
&\quad= \frac{(y-x)\mathrm{e}^{\mathrm{e}y-x-g(n)}}{\sqrt{2\pi\mathrm{e}}}. \qquad (2.17)
\end{aligned}
$$

We have

$$
\lim_{n\to\infty}\frac{(y-x)\mathrm{e}^{\mathrm{e}y-x-g(n)}}{\sqrt{2\pi\mathrm{e}}}=0,
$$

giving us our result.

For Theorem 1.2(iii), again let $x<y$ be constants. We will show that

$$
\lim_{n\to\infty}\mathbb{P}(N(n,nx)\le T_m(n)\le N(n,ny),\, S_k=0)=0,
$$

which implies the desired result since $x<y$ are arbitrary. We have

$$
\begin{aligned}
\mathbb{P}(N(n,nx)\le T_m(n)\le N(n,ny),\, S_k=0) &\le \sum_{M=\lceil N(n,nx)\rceil}^{\lfloor N(n,ny)\rfloor}\frac{\mathbb{P}(S_{m-1,M-1}=1,\, S_{k,M-1}=0)}{n} \\
&\le \sum_{M=\lceil N(n,nx)\rceil}^{\lfloor N(n,ny)\rfloor}\frac{\mathbb{P}(S_{k,M-1}=0)}{n}. \qquad (2.18)
\end{aligned}
$$

https://doi.org/10.1017/jpr.2022.94 Published online by Cambridge University Press

Note that

$$\mathbb{E}(S_{k,M}) = \binom{M}{k}\left(1-\frac{1}{n}\right)^{M-k}\frac{1}{n^{k-1}},$$

$$\mathbb{E}(S_{k,M}^2) = \binom{M}{k}\binom{M-k}{k}\left(1-\frac{2}{n}\right)^{M-2k}\frac{(n-1)}{n^{2k-1}} + \binom{M}{k}\left(1-\frac{1}{n}\right)^{M-k}\frac{1}{n^{k-1}}$$

$$< \binom{M}{k}^2\left(1-\frac{1}{n}\right)^{2M-4k}\frac{1}{n^{2k-2}} + \binom{M}{k}\left(1-\frac{1}{n}\right)^{M-k}\frac{1}{n^{k-1}}.$$

We have

$$\mathrm{Var}(S_{k,M}) = \mathbb{E}(S_{k,M}^2) - \mathbb{E}(S_{k,M})^2 \le \mathbb{E}(S_{k,M})^2\left(\left(1-\frac{1}{n}\right)^{-2k} - 1\right) + \mathbb{E}(S_{k,M}).$$

Since $k < n$ for sufficiently large n, we have that there exists $C_1 > 0$ such that, for sufficiently large n,

$$\mathrm{Var}(S_{k,M}) \le \frac{C_1 k\mathbb{E}(S_{k,M})^2}{n} + \mathbb{E}(S_{k,M}).$$

Similarly to the derivation of (2.17), we can deduce that there exists $C_2 > 0$ depending only on x and y such that, for sufficiently large n, $\mathbb{E}(S_{k,M}) < C_2 \mathrm{e}^{g(n)} < C_2 \log n$. Therefore, for sufficiently large n, we have $\mathrm{Var}(S_{k,M}) \le \mathbb{E}(S_{k,M}) + 1 < 2\mathbb{E}(S_{k,M})$. Thus, for sufficiently large n, $\sigma(S_{k,M}) < \sqrt{2\mathbb{E}(S_{k,M})}$. By Chebyshev's inequality, we therefore have

$$\begin{aligned}\mathbb{P}(S_{k,M} = 0) &\le \mathbb{P}(|S_{k,M} - \mathbb{E}(S_{k,M})| \ge \mathbb{E}(S_{k,M})) \\ &\le \mathbb{P}\left(|S_{k,M} - \mathbb{E}(S_{k,M})| \ge \frac{\sigma(S_{k,M})\sqrt{\mathbb{E}(S_{k,M})}}{\sqrt{2}}\right) \le \frac{2}{\mathbb{E}(S_{k,M})} \qquad (2.19)\end{aligned}$$

for sufficiently large n. Also, similarly to the derivation of (2.17), we have that there exists $C_3 > 0$ depending on only x and y such that $C_3 \mathrm{e}^{g(n)} < \mathbb{E}(S_{k,M})$. Thus, from (2.18) and (2.19), we have

$$\mathbb{P}(N(n, nx) \le T_m(n) \le N(n, ny), S_k = 0) \le \sum_{M=\lceil N(n,nx)\rceil}^{\lfloor N(n,ny)\rfloor} \frac{2}{C_3 n \mathrm{e}^{g(n)}} = \frac{2(y-x)}{C_3 \mathrm{e}^{g(n)}}$$

for sufficiently large n. Since $\lim_{n\to\infty} g(n) = \infty$, we have our result. □

3. Future work

There are a few open questions that are worth exploring with respect to the coupon collector's problem and the number of k-tons. For instance, can we extend our ranges for k? We have results for $k = o(\log n)$ if $m = 1$ and $k = o(\log n/\log\log n)$ if $m \ge 2$, as well as $k = \mathrm{e}\log n + \left((\mathrm{e}-1)(m-1) - \frac{1}{2}\right)\log\log n + d$, but we can also ask similar questions for values of k between these minimum and maximum values. For instance, we can consider $k = \theta(\log n)$ or $k = \theta(\log n/\log\log n)$. We can also ask what happens if we have m increasing with n. Also, [3] studied the limiting distribution of $T_m(n)$ in the case of unequal coupon probabilities. We could also study the limiting distribution of k-tons in the case of unequal coupon probabilities.

https://doi.org/10.1017/jpr.2022.94 Published online by Cambridge University Press

Acknowledgement

The author would like to thank Dr. Daniel Berend for helpful comments on the paper.

Funding information

The research of J. C. Saunders is supported by an Azrieli International Postdoctoral Fellowship, as well as a Postdoctoral Fellowship at the University of Calgary.

Competing interests

There were no competing interests to declare which arose during the preparation or publication process of this article.

References

[1] BAUM, L. AND BILLINGSLEY, P. (1965). Asymptotic distributions for the coupon collector's problem. *Ann. Math. Statist.* **36**, 1835–1839.
[2] BERENBRINK, P. AND SAUERWALD, T. (2009). The weighted coupon collector's problem and applications. In *Computing and Combinatorics* (LNCS 5609), ed H. Q. Ngo. Springer, Berlin, pp. 449–458.
[3] DOUMAS, A. AND PAPANICOLAOU, V. (2012). The coupon collector's problem revisited: Asymptotics of the variance. *Adv. Appl. Prob.* **44**, 166–195.
[4] DOUMAS, A. AND PAPANICOLAOU, V. (2016). The coupon collector's problem revisited: Generalizing the double Dixie cup problem of Newman and Shepp. *ESIAM Prob. Statist.* **20**, 367–399.
[5] ERDÖS, P. AND RÉNYI, A. (1961). On a classical problem of probability theory. *Magyar Tudományos Akadémia Matematikai Kutató Intézetének Közleményei* **6**, 215–220.
[6] FOATA, D. AND HAN, G. (2001). Les nombres hyperharmoniques et la fratrie du collectionneur de vignettes. *Séminaire Lotharingien de Combinatoire* **47**, B47a.
[7] MYERS, A. AND WILF, H. (2006). Some new aspects of the coupon collector's problem. *SIAM Rev.* **48**, 549–565.
[8] NEAL, P. (2008). The generalised coupon collector problem. *J. Appl. Prob.* **45**, 621–629.
[9] NEWMAN, D. AND SHEPP, L. (1960). The double Dixie cup problem. *Amer. Math. Monthly* **67**, 58–61.
[10] PENROSE, M. (2009). Normal approximation for isolated balls in an urn allocation model. *Electron. J. Prob.* **14**, 2155–2181.