The Generalized Coupon Collector Problem

John Moriarty & Peter Neal

First version: 30 April 2008

Research Report No. 9, 2008, Probability and Statistics Group
School of Mathematics, The University of Manchester

Applied Probability Trust (25 April 2008)

THE GENERALIZED COUPON COLLECTOR PROBLEM

JOHN MORIARTY,* *University of Manchester*

PETER NEAL,* *University of Manchester*

Abstract

Coupons are collected one at a time from a population containing n distinct types of coupon. The process is repeated until all n coupon have been collected and the total number of draws, Y, from the population is recorded. It is assumed that the draws from the population are independent and identically distributed (draws with replacement) according to a probability distribution X with the probability that a type i coupon is drawn being $P(X = i)$. The special case where each type of coupon is equally likely to be drawn from the population is the classic coupon collector problem. We consider the asymptotic distribution Y (appropriately normalized) as the number of coupons $n \to \infty$ under general assumptions upon the asymptotic distribution of X. The results are proved by studying the total number of coupons, $W(t)$, not collected in t draws from the population and noting that $\mathbb{P}(Y \leq t) = \mathbb{P}(W(t) = 0)$. Two normalizations of Y are considered, the choice of normalization depending upon whether or not a suitable Poisson limit exists for $W(t)$. Finally, extensions to the K-coupon collector problem and the birthday problem are given.

Keywords: The coupon collector problem; Poisson convergence; birthday problem.

2000 Mathematics Subject Classification: Primary 60F05

Secondary 60G70

1. Introduction

The classic coupon collector problem has a long history, see for example [3]. The classic problem is as follows. A collector wishes to collect a complete set of n distinct coupons, labeled 1 through to n. The coupons are hidden inside breakfast cereal boxes

* Postal address: School of Mathematics, University of Manchester, Alan Turing Building, Oxford Rd, Manchester, M13 9PL, United Kingdom

and within each cereal box there is one coupon which is equally likely to be any of the n distinct coupons. The collector purchases one box of breakfast cereals at a time, collecting the coupons, stopping when the collector has completed the set of n distinct coupons. The total number of cereal boxes, Y_n, which the collector needs to purchase is the quantity of interest. Elementary calculations show that

$$\mathbb{E}[Y_n] = n \sum_{i=1}^{n} \frac{1}{i} \approx n \log n.$$

Furthermore, if Z is a standard Gumbel distribution with $P(Z \leq z) = \exp(-\exp(-z))$ $(z \in \mathbb{R})$, then

$$\frac{1}{n}(Y_n - n \log n) \xrightarrow{D} Z \qquad \text{as } n \to \infty,$$

where '$\xrightarrow{D}$' denotes convergence in distribution, see for example [4].

The generalized coupon collector problem assumes that whilst the cereal boxes are independent and identically distributed, the probability that a box contains coupon i is p_i. No assumption is placed upon the $\{p_i\}$'s except that $p_i > 0$ $(i = 1, 2, \ldots, n)$. We allow for the possibility that some boxes may not contain a coupon by only assuming that $\sum_{i=1}^n p_i \leq 1$. The random coupon collector problem, [5, 4], is an alternative departure from the classic problem. The proofs in [4] rely upon a Poisson embedding argument and although our proofs are different we shall also exploit a Poisson approximation approach.

The paper is structured as follows. In Section 2 the main result, Theorem 2.1 is presented and proved. An alternative result is given in theorem 2.2 which is applicable when the Poisson arguments of theorem 2.1 fail. A number of examples are considered in section 3. Finally, in Section 4 extensions of Section 2 are discussed. These include the K-coupon collector problem, the total number of draws from the population that are required to have K coupons of each type and the K-birthday problem, the total number of draws from the population that are required to have K coupons of any (unspecified) type.

2. Coupon Collecting problem

For the asymptotic results of this paper, we consider a sequence of coupon collections $\{\mathcal{C}_n\}$ where the number of coupons to be collected $n \to \infty$. For $n \geq 1$, $\mathcal{C}_n$ requires the collection of n coupons, labeled 1 through to n to collect. Coupons are collected as follows. Let $X_1^n, X_2^n, \dots$ be independent and identically distributed according to X^n, where

$$\mathbb{P}(X^n = i) = \begin{cases} p_{ni} & i = 1, 2, \dots, n, \\ 0 & \text{otherwise}, \end{cases}$$

where $\sum_{i=1}^n p_{ni} \leq 1$ and $\min_{1 \leq i \leq n} p_{ni} > 0$. Then X_k^n is the k^{th} coupon drawn from the population (of coupons) and the process is continued until all n coupons have been collected. Let Y_n denote the total number of coupons which need to be collected to obtain the full set of coupons in $\mathcal{C}_n$.

Before stating the main result we introduce some useful notation. For $n \geq 1$, $i = 1, 2 \dots, n$ and $t = 1, 2, \dots$, let $\chi_i^n(t) = 1$ if coupon has not been collected in the first t coupons drawn from the population and $\chi_i^n(t) = 0$ otherwise. Let $W_n(t) = \sum_{i=1}^n \chi_i^n(t)$, the total number of distinct coupons which still need to be collected after t coupon draws. Thus for $t \geq 1$, $Y_n \leq t$ if and only if $W_n(t) = 0$.

Theorem 2.1. *Suppose that there exists sequences $\{b_n\}$ and $\{k_n\}$ such that $k_n/b_n \to 0$ as $n \to \infty$ and that for $y \in \mathbb{R}$,*

$$\sum_{i=1}^n \exp(-p_{ni}\{b_n + y k_n\}) \to g(y) \qquad \text{as } n \to \infty, \tag{2.1}$$

for a non-increasing function $g(\cdot)$ with $g(y) \to \infty$ as $y \to -\infty$ and $g(y) \to 0$ as $y \to 0$.

Then if $\tilde{Y}_n = (Y_n - b_n)/k_n$,

$$\tilde{Y}_n \xrightarrow{D} Y \qquad \text{as } n \to \infty,$$

where Y has cumulative distribution function

$$\mathbb{P}(Y \leq y) = \exp(-g(y)) \qquad (y \in \mathbb{R}).$$

The key restriction in Theorem 2.1 is that (2.1), implies that $\min_{1 \leq i \leq n} p_{ni} b_n \to \infty$ as $n \to \infty$. This condition is needed for the Poisson limit (2.3) below since it implies that $\max_{1 \leq i \leq n} \mathbb{E}[\chi_i^n([b_n + y k_n])] \to 0$ as $n \to \infty$. In Theorem 2.2 we explore the case where $\min_{1 \leq i \leq n} p_{ni} b_n \to c$ as $n \to \infty$, for some $0 < c < \infty$. By Jensen's inequality,

$$\begin{aligned} \sum_{i=1}^n \exp(-p_{ni} b_n) &\geq \sum_{i=1}^n \exp\left(-\frac{1}{n} b_n\right) \\ &= n \exp\left(-\frac{b_n}{n}\right). \end{aligned} \tag{2.2}$$

Therefore $b_n \geq n \log n$ and this will be used in Lemma 2.2. The only restriction placed upon the sequence $\{X^n\}$ is (2.1). Discussion of a natural construction of suitable sequences $\{X^n\}$ is deferred to Section 3.

The proof of theorem 2.1 relies upon two preliminary lemmas which are motivated and proved in the following discussion.

Since for $t \geq 1$, $Y_n \leq t$ if and only $W_n(t) = 0$, it suffices to show that for all $y \in \mathbb{R}$,

$$W_n([b_n + yk_n]) \xrightarrow{D} Po(g(y)) \qquad (y \in \mathbb{R}). \tag{2.3}$$

The first step in proving (2.3) is to show that for any $t \in \mathbb{N}$, $\{\chi_i^n(t)\}$ are negatively related, [1], page 24. For $n, t \geq 1$ and $1 \leq j \leq n$, let $\{\theta_{i,j}^n(t); i = 1, 2, \ldots, n\}$ be random variables satisfying

$$\mathcal{L}(\theta_{i,j}^n(t); i = 1, 2, \ldots, n) = \mathcal{L}(\chi_i^n(t); i = 1, 2, \ldots, n | \chi_j^n(t) = 1).$$

Lemma 2.1. *For $n, t \geq 1$, the random variables $\{\chi_i^n(t)\}$ are negatively related, i.e. for each $1 \leq j \leq n$, the random variables $\{\theta_{i,j}^n(t); i = 1, 2, \ldots, n\}$ and $\{\chi_i^n(t); i = 1, 2, \ldots, n\}$ can be defined on a common probability space $(\Omega, \mathcal{F}, P)$ such that, for all $i \neq j$, $\chi_i^n(t)(\omega) \geq \theta_{i,j}^n(t)(\omega)$ for all $\omega \in \Omega$.*

Proof. The lemma is proved by a simple coupling argument.

Fix $n, t \geq 1$ and $j = 1, 2, \ldots, n$. Draw $X_1^n, X_2^n, \ldots, X_t^n$ from X^n. For $k = 1, 2, \ldots, t$, let $\tilde{X}_k^n(t) \stackrel{D}{=} X_k^n | \chi_j^n(t) = 1$. For $k = 1, 2, \ldots, t$, if $X_k^n \neq j$, set $\tilde{X}_k^n(t) = X_k^n$. If $X_k^n = j$, set $\tilde{X}_k^n(t) = \hat{X}_k^n$, where

$$\mathbb{P}(\hat{X}_k^n = i) = \begin{cases} \frac{p_{ni}}{1 - p_{nj}} & i \neq j \\ 0 & \text{otherwise.} \end{cases}$$

Thus $\tilde{X}_1^n(t), \tilde{X}_2^n(t), \ldots, \tilde{X}_t^n(t)$ have the correct distribution and by construction $\chi_i^n(t) \geq \theta_{i,j}^n(t)$ for $i \neq j$.

Note that

$$\begin{aligned} \mathbb{E}[W_n([b_n + yk_n])] &= \sum_{i=1}^n (1 - p_{ni})^{[b_n + yk_n]} \\ &\to g(y) \qquad \text{as } n \to \infty. \end{aligned}$$

Therefore by Lemma 2.1 and [1], Corollary 2.C.2, (2.3) holds if

$$var(W_n([b_n + yk_n]) \to g(y) \qquad \text{as } n \to \infty. \tag{2.4}$$

Now $var(W_n([b_n + yk_n])$ is equal to

$$\sum_{i=1}^{n} var(\chi_i^n([b_n + yk_n])) + \sum_{i=1}^{n}\sum_{j\neq i} cov(\chi_i^n([b_n + yk_n]), \chi_j^n([b_n + yk_n])). \tag{2.5}$$

Equation (2.1) ensures that

$$\sum_{i=1}^{n} \exp(-p_{ni}[b_n + yk_n])^2 \to 0 \quad \text{as } n \to \infty.$$

Therefore by (2.1), the first term in (2.5) converges to $g(y)$ as $n \to \infty$. Thus (2.4) holds if the latter term in (2.5) converges to 0 as $n \to \infty$.

Lemma 2.2.

$$\sum_{i=1}^{n}\sum_{j\neq i} |cov(\chi_i^n([b_n + yk_n]), \chi_j^n([b_n + yk_n]))| \to 0 \quad \textit{as } n \to \infty.$$

Proof. For any $i \neq j$,

$$\begin{aligned} & |cov(\chi_i^n([b_n + yk_n]), \chi_j^n([b_n + yk_n]))| \\ = & \left|(1 - p_{ni} - p_{nj})^{[b_n+yk_n]} - (1 - p_{ni})^{[b_n+yk_n]}(1 - p_{nj})^{[b_n+yk_n]}\right| \\ = & (1 - p_{ni})^{[b_n+yk_n]}(1 - p_{nj})^{[b_n+yk_n]} \left|\left(1 - \frac{p_{ni}p_{nj}}{(1 - p_{ni})(1 - p_{nj})}\right)^{[b_n+yk_n]} - 1\right| \\ \leq & (1 - p_{ni})^{[bn \log n + yn]}(1 - p_{nj})^{[b_n+yk_n]} \times \frac{[b_n + yk_n]p_{ni}p_{nj}}{(1 - p_{ni})(1 - p_{nj})}, \end{aligned}$$

with the inequality coming from $|1 - (1 - y)^m| \leq my$ for $0 \leq y \leq 1$ and $m \in \mathbb{N}$.

Therefore

$$\begin{aligned} & \sum_{i=1}^{n}\sum_{j\neq i} |cov(\chi_i^n([b_n + yk_n]), \chi_j^n([b_n + yk_n]))| \\ \leq & \left\{\sqrt{[b_n + yk_n]} \sum_{i=1}^{n} \frac{p_{ni}}{1 - p_{ni}}(1 - p_{ni})^{[b_n+yk_n]}\right\}^2. \end{aligned} \tag{2.6}$$

Let $\mathcal{A}_n = \{i; p_{ni} \leq b_n^{-3/4}\}$. Then

$$\begin{aligned} & \sqrt{[b_n + yk_n]} \sum_{i=1}^{n} \frac{p_{ni}}{1 - p_{ni}}(1 - p_{ni})^{[b_n+yk_n]} \\ = & \sqrt{[b_n + yk_n]} \sum_{i\in\mathcal{A}_n} \frac{p_{ni}}{1 - p_{ni}}(1 - p_{ni})^{[b_n+yk_n]} + \sqrt{[b_n + yk_n]} \sum_{i\in\mathcal{A}_n^C} \frac{p_{ni}}{1 - p_{ni}}(1 - p_{ni})^{[b_n+yk_n]} \\ \leq & \frac{b_n^{-3/4}\sqrt{[b_n + yk_n]}}{1 - b_n^{-3/4}} \sum_{i=1}^{n}(1 - p_{ni})^{[b_n+yk_n]} + \sqrt{[b_n + yk_n]} \sum_{i\in\mathcal{A}_n^C} (1 - b_n^{-3/4})^{[b_n+yk_n]-1} \\ \to & 0 \quad \text{as } n \to \infty, \end{aligned}$$

since $\sum_{i=1}^n (1-p_{ni})^{[b_n+yk_n]} \to g(y)$ and $b_n \geq n \log n$ as $n \to \infty$. Therefore the right-hand side of (2.6) converges to 0 as $n \to \infty$ and the lemma is proved.

Proof of Theorem 2.1. For any $y \in \mathbb{R}$, $\tilde{Y}_n \leq y$ if and only if $W_n([b_n + yk_n]) = 0$. Therefore by (2.3), for $y \in \mathbb{R}$,

$$\begin{aligned} \mathbb{P}(\tilde{Y}_n \leq y) &= \mathbb{P}(W_n([b_n + yk_n]) = 0) \\ &\to \exp(-g(y)) = \mathbb{P}(Y \leq y) \quad \text{as } n \to \infty, \end{aligned}$$

and the theorem is proved.

The proof of theorem 2.1 presents a straightforward bound for $|\mathbb{P}(\tilde{Y}_n \leq y) - \mathbb{P}(Y \leq y)|$ $(y \in \mathbb{R})$. For $t \geq 0$, let $Z(t) \sim Po(t)$ and for $y \in \mathbb{R}$, let $g_n(y) = \mathbb{E}[W_n([b_n + yk_n])]$. By the triangle inequality and [1], Corollary 2.C.2,

$$\begin{aligned} & |\mathbb{P}(\tilde{Y}_n \leq y) - \mathbb{P}(Y \leq y)| \\ = \ & |\mathbb{P}(W_n([b_n + yk_n]) = 0) - \mathbb{P}(Z(g(y)) = 0)| \\ \leq \ & |\mathbb{P}(W_n([b_n + yk_n]) = 0) - \mathbb{P}(Z(g_n(y)) = 0)| + |\mathbb{P}(Z(g_n(y)) = 0) - \mathbb{P}(Z(g(y)) = 0)| \\ \leq \ & \left(1 - e^{-g_n(y)}\right)\left(1 - \frac{var(W_n([b_n + yk_n]))}{g_n(y)}\right) + \left|e^{-g_n(y)} - e^{-g(y)}\right|. \end{aligned}$$

We now turn our attention to the situation where the natural scaling $\{b_n\}$ is such that $\min_{1 \leq i \leq n} p_{ni} b_n \to c$ as $n \to \infty$, for some $0 < c < \infty$.

Theorem 2.2. *Suppose that there exists sequences $\{b_n\}$ such that for $y \in \mathbb{R}^+$,*

$$\sum_{i=1}^n \exp(-p_{ni} y b_n) \to g(y) \quad \text{as } n \to \infty, \tag{2.7}$$

for a non-increasing function $g(\cdot)$ with $g(y) \to \infty$ as $y \to 0$ and $g(y) \to 0$ as $y \to \infty$.

Suppose that there exists a function $h(\cdot)$ such that for all $y \in \mathbb{R}^+$,

$$\prod_{i=1}^n (1 - \exp(-p_{ni} y b_n)) \to h(y) \quad \text{as } n \to \infty. \tag{2.8}$$

Then (2.7) ensures that $h(y) \to 0$ as $y \to 0$ and $h(y) \to 1$ as $y \to \infty$, and if $\hat{Y}_n = Y_n / b_n$,

$$\hat{Y}_n \xrightarrow{D} Y \quad \text{as } n \to \infty,$$

where Y has cumulative distribution function

$$\mathbb{P}(Y \leq y) = h(y) \quad (y \in \mathbb{R}^+).$$

Proof. The proof has a number of similarities and differences to the proof of theorem 2.1. We shall again exploit that $Y_n \leq t$ if and only $W_n(t) = 0$.

Let η_*^n be a homogeneous Poisson point process with rate 1 and let $T_n(t)$ denote the time of the $[tb_n]^{th}$ point on η_*^n. Let $V_1^n, V_2^n, \ldots$ be independent and identically distributed according to X^n. Let $\eta_1^n, \eta_2^n, \ldots, \eta_n^n$ be independent homogeneous Poisson point processes with rates $p_{n1}, p_{n2}, \ldots, p_{nn}$, respectively, constructed from η_*^n and $V_1^n, V_2^n, \ldots$ as follows. For $k = 1, 2, \ldots$, let s_k^n denote the time of the k^{th} point on η_*^n then there is a point on η_j^n at time s_k^n if $V_k^n = j$. Furthermore, $\chi_1^n(t), \chi_2^n(t), \ldots, \chi_n^n(t)$, and hence, $W_n(t)$ can be constructed using $V_1^n, V_2^n, \ldots, V_t^n$.

Let $\psi_i^n(t) = 1$ if there is no point on $\eta_i^n[0,t]$ and note that $\{\psi_i^n(t)\}$'s are independent. For $t \geq 0$, let $\tilde{W}_n(t) = \sum_{i=1}^n \psi_i^n(t)$. Then $W_n([yb_n]) = \tilde{W}_n(T_n([yb_n]))$. Since $\tilde{W}_n(\cdot)$ is non-decreasing, if $[yb_n] - ([yb_n])^{3/4} \leq T_n([yb_n]) \leq [yb_n] + ([yb_n])^{3/4}$ then

$$\tilde{W}_n([yb_n] + ([yb_n])^{3/4}) \leq W_n([yb_n]) \leq \tilde{W}_n([yb_n] - ([yb_n])^{3/4}). \tag{2.9}$$

Since $\frac{1}{(yb_n)^{3/4}}(T_n([yb_n]) - [yb_n]) \xrightarrow{p} 0$ as $n \to \infty$, it follows from (2.9) that $\mathbb{P}(W_n([yb_n]) = 0) \to h(y)$ if

$$\mathbb{P}(\tilde{W}_n([yb_n] \pm ([yb_n])^{3/4}) = 0) \quad \to \quad h(y) \qquad \text{as } n \to \infty.$$

By independence, for all $y \in \mathbb{R}$,

$$\begin{aligned} \mathbb{P}(\tilde{W}_n([yb_n] \pm ([yb_n])^{3/4}) = 0) &= \prod_{i=1}^n \left(1 - (1 - p_{ni})^{([yb_n] \pm ([yb_n])^{3/4})}\right) \\ &\to h(y) \qquad \text{as } n \to \infty. \end{aligned}$$

The main benefit of Theorem 2.1 over Theorem 2.2 is that $g(y)$ is usually much easier to calculate than $h(y)$.

3. Examples

A natural construction of $\{X^n\}$ is to take a (continuous) distribution X with probability density function $f(\cdot)$ on $[0,1]$ and for $n = 1, 2, \ldots$ and $i = 1, 2, \ldots, n$, set

$$p_{ni} = \int_{(i-1)/n}^{i/n} f(x)\, dx. \tag{3.1}$$

A number of results can be proved concerning various choices of X with Lemma 3.1 illustrating the point using a class of distributions with $f(\cdot)$ being continuous.

Lemma 3.1. *Let $0 \leq \sigma \leq 1$ be such that for all $0 \leq x \leq 1$ and $x \neq \sigma$, $0 < f(\sigma) < f(x)$. For $p = 1, 2$, let*

$$\begin{aligned} u_p &= \lim_{\epsilon \to 0+} \frac{f(\sigma+\epsilon) - f(\sigma)}{\epsilon^p} \\ l_p &= \lim_{\epsilon \to 0-} \frac{f(\sigma+\epsilon) - f(\sigma)}{|\epsilon|^p}. \end{aligned}$$

(*i*) *Suppose that $1_{\{\sigma>0\}} l_1 + 1_{\{\sigma<1\}} u_1 > 0$. Then $b_n = \frac{n}{f(\sigma)}(\log n - \log(\log n))$ and $k_n = n$ with*

$$g(y) = f(\sigma) \left(\frac{1_{\{\sigma>0\}}}{l_1} + \frac{1_{\{\sigma<1\}}}{u_1} \right) \exp(-f(\sigma) y).$$

(*ii*) *Suppose that $1_{\{\sigma>0\}} l_1 + 1_{\{\sigma<1\}} u_1 = 0$ and $1_{\{\sigma>0\}} l_2 + 1_{\{\sigma<1\}} u_2 > 0$. Then $b_n = \frac{n}{f(\sigma)} \left(\log n - \frac{1}{2} \log(\log n) \right)$ and $k_n = n$ with*

$$g(y) = \sqrt{\frac{\pi f(\sigma)}{2}} \left(\sqrt{\frac{1_{\{\sigma>0\}}}{l_2}} + \sqrt{\frac{1_{\{\sigma<1\}}}{u_2}} \right) \exp(-f(\sigma) y).$$

Proof. We outline the proof of (*i*) with (*ii*) being proved similarly.

Let $b_n = \frac{n}{f(\sigma)}(\log n - \log(\log n))$ and $k_n = n$. Note that

$$\begin{aligned} \sum_{i=1}^n \exp(-p_{ni}(b_n + y k_n)) &\approx \sum_{i=1}^n \exp\left(-(b_n + y k_n) \frac{1}{n} f\left(\frac{i - 1/2}{n} \right) \right) \\ &= n \sum_{i=1}^n \frac{1}{n} \exp\left(-\left(\frac{b_n}{n} + y \right) f\left(\frac{i-1/2}{n} \right) \right) \\ &\approx n \int_0^1 \exp\left(-\left(\frac{b_n}{n} + y \right) f(x) \right) dx. \end{aligned}$$

Therefore it is straightforward to show that

$$g(y) = \lim_{n \to \infty} n \int_0^1 \exp\left(-\left(\frac{1}{f(\sigma)} (\log n - \log(\log n)) + y \right) f(x) \right) dx.$$

Linearizing $f(x)$ about σ and considering the left and right hand limits separately yields the result.

Examples of probability density functions on $[0,1]$ satisfying Lemma 3.1 include $f(x) = \frac{2}{3}(1+x)$, $f(x) = \frac{6}{5}(1 - x(1-x))$ and $f(x) = \frac{12}{7}\max(1-x, x/2)$.

Suppose instead that X is piecewise constant with for $1 \leq j \leq k$,

$$f(x) = \lambda_j \qquad (\pi_{j-1} < x \leq \pi_j)$$

where $\lambda_1, \lambda_2, \ldots, \lambda_k > 0$ and $0 = \pi_0 < \pi_1 < \ldots < \pi_k = 1$. Without loss of generality assume that $\lambda_1 < \lambda_2 < \ldots < \lambda_k$. Then $b_n = \frac{1}{\lambda_1} n \log n$, $k_n = n$ and $g(y) = \pi_1 \exp(-\lambda_1 y)$.

In the above examples $k_n/b_n \to 0$ and Theorem 2.1 holds. In all cases, the limiting distribution Y is a Gumbel distribution with $b_n/n \log n \to 1/\min_{0 \leq x \leq 1} f(x)$ as $n \to \infty$.

An example of where Theorem 2.2 is necessary is $f(x) = 2x$ $(0 \leq x \leq 1)$ giving $p_{ni} = \frac{2i-1}{n^2}$ $(i = 1, 2, \ldots, n)$. Then for $y \in \mathbb{R}^+$,

$$\sum_{i=1}^{n} \exp(-p_{ni} y n^2) = \sum_{i=1}^{n} \exp(-(2i-1)y) \to g(y) = \frac{e^y}{e^{2y} - 1} \qquad \text{as } n \to \infty,$$

and theorem 2.2 holds with $b_n = n^2$ and $h(y) = \lim_{n \to \infty} \prod_{i=1}^{n} (1 - \exp(-(2i-1)y))$.

4. Extensions

The methodology outlined in Section 2 can be extended to find the total number of coupons, Y_n^K, which need to be collected in order to have (at least) K coupons of each type. In this case, simply let $\chi_i^n(t) = 1$ if at most $K-1$ coupons of type i have been collected in the first t draws from the population and $\chi_i^n(t) = 0$ otherwise. Then set $W_n^K(t) = \sum_{i=1}^n \chi_i^n(t)$ and note that $Y_n^K \leq t$ if and only if $W_n^K(t) = 0$. It is straightforward to adapt Lemmas 2.1 and 2.2 to this case and consequently, Theorem 2.1 holds with (2.1) replaced by

$$\frac{b_n^{K-1}}{(K-1)!} \sum_{i=1}^{n} p_{ni}^{K-1} \exp\left(-p_{ni}\{b_n + y k_n\}\right) \to g(y) \qquad \text{as } n \to \infty. \tag{4.1}$$

Since $k_n/b_n \to 0$ implies that $\min_{1 \leq i \leq n} b_n p_{ni} \to \infty$ as $n \to \infty$, (4.1) holds if and only if $\mathbb{E}[W_n^K([b_n + y k_n])] \to g(y)$ as $n \to \infty$. Theorem 2.2 can also be adapted to the K-coupon collector problem.

At the other end of the spectrum, the Poisson arguments above can be applied to the generalized birthday problem. That is, for $K \geq 2$, let U_n^K denote the total number of draws from the population that are required to obtain K coupons of any (unspecified) type. Let $\tilde{\chi}_i^n(t) = 1$ if at least K coupons of type i have been collected in the first t draws from the population and $\tilde{\chi}_i^n(t) = 0$ otherwise. Then if $\tilde{W}_n^K(t) = \sum_{i=1}^n \tilde{\chi}_i^n(t)$, $U_n^K > t$ if and only if $\tilde{W}_n^K(t) = 0$. Along the lines of Lemma 2.1 it can be shown that $\{\tilde{\chi}_i^n(t)\}$ are negatively related and straightforward bounds for the covariance terms can be obtained. We then have the following theorem.

Theorem 4.1. *For fixed $K \geq 2$, suppose that there exists a sequence $\{l_n\}$ such that*

$$l_n^K \sum_{i=1}^n p_{ni}^K \to 1, \tag{4.2}$$

and $\max_{1 \leq i \leq n} l_n p_{ni} \to 0$ *as $n \to \infty$. Then*

$$U_n^K / l_n \xrightarrow{D} U^K \qquad \text{as } n \to \infty,$$

where U^K has cumulative distribution function

$$\mathbb{P}(U^K \leq u) = 1 - \exp(-u^K) \qquad (u \in \mathbb{R}^+).$$

Proof. The conditions imposed on $\{l_n\}$ are sufficient for $W_n^K([ul_n]) \xrightarrow{D} Po(u^K)$ from which the theorem follows immediately.

The limiting distribution U^K obtained in Theorem 4.1 is identical to that obtained in [4], Theorem 5.2, for the random birthday problem. For the case $K = 2$, Theorem 4.1 follows immediately from [2], Example 2, since given (4.2), $\max_{1 \leq i \leq n} l_n p_{ni} \to 0$ if and only if $l_n^3 \sum_{i=1}^n p_{ni}^3 \to 0$ as $n \to \infty$.

Finally, it is worth noting that for the establishing of Poisson limits for $W_n^K([b_n + yk_n])$ and $\tilde{W}_n^K([ul_n])$ it is crucial that $\max_{1 \leq i \leq n} \mathbb{E}[\chi_i^n([b_n + yk_n])] \to 0$ and $\max_{1 \leq i \leq n} \mathbb{E}[\tilde{\chi}_i^n([ul_n])] \to 0$ as $n \to \infty$, respectively. That is, for the K-coupon collector problem we require that $\min_{1 \leq i \leq n} b_n p_{ni} \to \infty$ as $n \to \infty$ (none of the probabilities are too small) and for the K-birthday problem we require that $\max_{1 \leq i \leq n} l_n p_{ni} \to 0$ as $n \to \infty$ (none of the probabilities are too large).

References

[1] Barbour, A. D., Holst, L. and Janson, S. (1992). *Poisson Approximation.* Oxford University Press, Oxford.

[2] Blom, G. and Holst, L. (1989). Some properties of similar pairs. *Adv. Appl. Prob.*, **21**, 941–944.

[3] Feller, W. (1957). *An introduction to Probability Theory and its Applications.* Wiley, New York.

[4] Holst, L. (2001). Extreme value distributions for random coupon collector and birthday problems. *Extremes*, **4**, 129–145.

[5] Papanicolaou, V. G., Kokolakis, G. E. and Boneh, S. (1998) Asymptotics for the random coupon collector problem. *Journal of Computational and Applied Mathematics*, **93**, 95–105.