Revised version of *Stochastic Models*, **13**, #1 39–66 (1997).

THE COUPON-COLLECTOR PROBLEM REVISITED — A SURVEY OF ENGINEERING PROBLEMS AND COMPUTATIONAL METHODS

Arnon Boneh
School of Business Administration
Tel-Aviv University
Ramat Aviv, Tel-Aviv, Israel
boneh@techunix.technion.ac.il

Micha Hofri
Department of Computer Science
M/S 132 Rice University
6100 Main St. Houston, TX 77005-1892
hofri@cs.rice.edu

ABSTRACT

A standard combinatorial problem is to estimate the number (T) of coupons, drawn at random, needed to complete a collection of all possible m types. Generalizations of this problem have found many engineering applications. The usefulness of the model is hampered by the difficulties in obtaining numerical results for moments or distributions. We show two computational paradigms that are well suited for this type of problems: one, following Flajolet *et al.* [21], is the calculus of generating functions over regular languages. We use it to provide relatively efficient answers to several questions about the sampling process – we show it is possible to compute arbitrarily accurate approximations for quantities such as $E[T]$, in a time which is linear in m, for any type distribution, while an exponential time is required for exact calculation. It also leads to a proof of a long-standing folk-theorem, concerning the extremality of uniform reference probabilities. The second method is a generalization of the Poisson transform [25], which we use to discuss statistical estimation procedures.

Key words: coupon collector, engineering applications, combinatorics of languages, numerical methods, Poisson transform, prediction function.

1. INTRODUCTION

The Coupon-Collector Problem (CCP) is defined as follows: A set of m distinct objects (coupons, items...) is sampled with replacement by a collector. Each drawing produces the ith item with a fixed probability p_i, independently of all past events. Several variables are defined for the sequence of trials; all depend on the sample probability vector $\underline{\boldsymbol{p}} = (p_1, p_2, \cdots, p_m)$:
$X_n(\underline{\boldsymbol{p}})$ — The item number (or 'name') sampled on the n-th drawing.
$T(\underline{\boldsymbol{p}})$ — The number of the drawing that completes the collection.
$T_j(\underline{\boldsymbol{p}})$ — The number of drawings needed to complete a sub-collection of size j.
$T_C(\underline{\boldsymbol{p}})$ — The number of drawings to complete a specified sub-collection C.
$Y_n(\underline{\boldsymbol{p}})$ — The number of different items observed in the first n drawings.
$N_k(\underline{\boldsymbol{p}}, n)$ — The number of different items observed *exactly* k times during the first n trials.

Note: The verb 'complete' has here the meaning of 'complete for the first time'. The traditional symbols E and V are used to denote the expectation and variance of random variables.

The CCP has a huge literature. We indicate below the references relevant to our discussion. It is a natural framework for many combinatorial questions (see [19] and [35]), and correspondingly numerous applications – though these rarely preserve the simplicity of the original CCP. Here are a few engineering examples that generate CCP-like processes.

Applications of the CCP

(1) Detection of all necessary (also called 'hard' or non-redundant) constraints in a constrained optimization problem.

Boneh and Golan [6] describe PREDUCE (for Probabilistic REDUCE), the implementation of a class of algorithms for detecting such constraints, when the feasibility region is convex and of full dimension d. Each iteration of PREDUCE consists of selecting an interior feasible point, generating from it a ray in random direction, and recording the first constraint(s) it meets when extended. Such a constraint belongs to the hard-constraints set. The algorithm generates rays until a stopping rule is satisfied. All the constraints not hit by that time are assumed redundant – possibly erroneously. Each iteration corresponds to drawing one coupon. The number of items is known, but not the probabilities p_i, and they are not constant, because the selected interior points vary over the trials. Each p_i is proportional to the expected $d-1$ dimensional angle subtended by the corresponding facet, over the set of ray origins (see [43]). A direct determination of the hard constraints set is computationally very hard.

(2) Multistart [46]. This is a component of optimization algorithms, for both constrained and unconstrained environments with multiple local optima. It uses a deterministic algorithm, which, given an initial point, finds a local optimum of the objective function (the determinism means that repeating the starting point leads to the same local optimum). The algorithm thus partitions the search space into attraction regions, one for each local optimum of the objective function. If the initial point is chosen at random, the (unknown) probability of hitting a specific local optimum is proportional to the size of its attraction region. Here, in addition to the CCP-like question of stopping times for the sampling of initial points, one asks about the likelihood of obtaining a higher-valued local optimum.

(3) Determining the convex closure of a set of points $S \in R^n$.
This problem appears to be related to PREDUCE, but behaves quite differently. To determine the subset of S that spans the closure we generate random $n-1$-dimensional hyperplanes and compute the distances of all the points of S from each. The largest (absolute) values—one, if they are all of the same sign, or two otherwise—belong to points that are in the spanning subset. If there are closure hyperplanes that contain

more than $n-1$ points, some of these points will never be “discovered,” although they belong to the spanning set (visualize this for $n=2$ with triplets of collinear points: the intermediate ones have zero probability of being the farthest from any random line).

(4) The fault-detection (FD) problem in combinatorial circuits.
A combinatorial circuit can be viewed as a black box with two sets of pins: one for input and one for output. Each pin carries one bit (0 or 1) at any specified time. Loading the input pins with a bit configuration produces an output vector on the output set – which should agree with the specifications of the circuit. However, circuits fail sometimes. The standard fault model, “stuck at” [18], assumes the following:

(a) The only possible faults are of lines in the circuit that are stuck (at 0 or 1) independently of the input vector.

(b) Faults are rare events; the probability of having more than one fault at any one time—assuming no previous faults—is negligible. (This assumption is used to bound the duration of the test).

(c) Faults occur as independent events.

Furthermore, we assume that a list of all possible faults is available to the test designer.

A possible way to detect a fault is to find that an input vector produced an output which is in error by at least one bit. Correct output may be produced (for certain input vectors) by a faulty circuit. The FD problem consists of constructing a *minimal* list of input vectors—since in critical applications the test is performed frequently—which detect all the entries in the fault-list.

One approach to the FD problem is to select random input vectors, determine the faults detected by each input, and continue until all faults are ‘covered.’ Let n be the size of the input-pins set. If all n-long input vectors are generated equi-probably, then p_i, the probability of detecting the i-th fault, is $2^{-n} \times$ #(of input vectors producing wrong output when fault i is present). Clearly the search for a covering set is CCP-like: input vectors ‘draw’ faults at random. However, most vectors detect more than one fault, possibly hundreds. This may be represented within the CCP formalism either by saying that $\Sigma\, p_i > 1$, or by viewing each vector as corresponding to a batch of drawings of a random size. The distribution of the batch size can be estimated well by the designer. Further details may be found in [3] and [8].

(5) Testing biological cultures for contamination.
The last example is from Gail *et al.* [22], to which we refer the reader for the relevant biology. It differs from the other examples in that both the probability vector $\underline{\boldsymbol{p}}$ and m are known. The test compares the likelihood, which they compute, that no repetition occurs in n drawings, with experiments on potentially contaminated cells. While the same mathematics is used there as we do below, this is more similar to another classic – the ‘birthday problem’.

Outline of the paper

Section 2 surveys the main results in the literature and comments on the numerical difficulties which have stymied applications of the CCP model.

Section 3 describes our first computational device—operations on regular languages—and shows its effectiveness in producing efficiently computable expressions for moments of variables associated with the CCP. It also leads to a proof that among all possible $\underline{\boldsymbol{p}}$, the uniform vector produces the shortest expected collection completion times.

Section 4 describes numerical examples, and the computational techniques we used. Then we discuss statistical problems that arise when CCP-like processes are considered.

In Section 5 we introduce our second computational method, the Poisson transform, which leads to a new form of the prediction function.

2. RELATED WORK

Many texts on probability use the CCP as an example for elementary derivations of expectations. Feller [19] also considers the waiting times between successive increases of the "observed set." David and Barton consider in [16] the CCP within their discussion of occupancy problems, and compute the time required to fill a given number of boxes, remarking that the 'moments are not tractable' (though it was not computational complexity that seems to have concerned them, but rather the lack of closed forms for the results). Most of the treatments (we know) consider the expected time to complete the collection, $E[T(\underline{\boldsymbol{p}})]$, and its properties, usually in the classical case of equally likely (EL) coupons.

The best-known expression for $E[T(\underline{\boldsymbol{p}})]$ is probably

$$\begin{aligned} E[T(\underline{\boldsymbol{p}})] &= \sum_{i=1}^{m} \frac{1}{p_i} - \sum_{1\le i<j\le m} \frac{1}{p_i+p_j} + \sum_{1\le i<j<k\le m} \frac{1}{p_i+p_j+p_k} - + \cdots \\ &= \sum_{r=1}^{m} (-1)^{r+1} \sum_{1\le i_1<\cdots<i_r\le m} \frac{1}{p_{i_1}+p_{i_2}+\cdots+p_{i_r}}, \end{aligned} \tag{1}$$

easily proved [13] with the inclusion-exclusion principle and the relation $E[T(\underline{\boldsymbol{p}})] = \Sigma_{n\ge 0}\Pr[T(\underline{\boldsymbol{p}}) > n]$. The earliest source we found for it is [16], which provides some distributional results as well.

For the EL case, when $\underline{\boldsymbol{p}} = \underline{\boldsymbol{e}}/m$, with $\underline{\boldsymbol{e}} = (1, 1, \cdots, 1)$, it is possible to obtain compact expressions also for higher moments, since $T(\underline{\boldsymbol{e}}/m)$ is the sum of independent geometrically distributed random variables, with parameters that depend on their position in the sum (but not on the particular items that had been sampled):

$$E[T(\frac{\underline{\boldsymbol{e}}}{m})] = mH_m, \qquad V[T(\frac{\underline{\boldsymbol{e}}}{m})] = m^2 H^{(2)}_{m-1} - mH_{m-1}, \qquad E[z^{T(\underline{\boldsymbol{e}}/m)}] = z\prod_{j=1}^{m-1} \frac{(m-j)z}{m-jz}, \tag{2}$$

where H_m is the mth Harmonic number, and $H^{(2)}_m$ is the mth second-order Harmonic number, given by $\Sigma_{i=1}^{m} 1/i^2$; these converge rapidly to $\zeta(2) = \pi^2/6$. In this egalitarian case even the PMF has a concise representation [28, p.129], in terms of the Stirling numbers of the second kind:

$$\Pr\left[T(\underline{\boldsymbol{e}}/m) = n\right] = \frac{m!}{m^n} \left\langle \begin{matrix} n-1 \\ m-1 \end{matrix} \right\rangle. \tag{3}$$

The expression for $E[T(\underline{\boldsymbol{p}})]$ above is a sum of a large number of terms with alternating signs, which is difficult to handle numerically; most approximation schemes are defeated in such circumstances as well. Other expressions were developed to overcome this difficulty, but they are even more complex.

A. Boneh obtains in [7] an expression for $E[T(\underline{\boldsymbol{p}})]$ by considering all possible orders in which the coupons may be obtained, and conditioning on the permutation. He finds

$$E[T(\underline{\boldsymbol{p}})] = \sum_{\underline{\boldsymbol{i}} \in S(N_m)} P(\underline{\boldsymbol{i}})E(\underline{\boldsymbol{i}}), \tag{4}$$

where N_m is the set of the first m natural numbers and $S(N_m)$ is the symmetric group of permutations of N_m. For any permutation $\underline{\boldsymbol{i}} = (i_1, \cdots, i_m)$ he writes

$$P(\underline{\boldsymbol{i}}) = \frac{\prod_{r=1}^{m} p_r}{\prod_{j=1}^{m}\Big(\sum_{k=j}^{m} p_{i_k}\Big)}, \qquad E(\underline{\boldsymbol{i}}) = \sum_{r=0}^{m-1} \frac{1}{1 - \sum_{k=1}^{r} p_{i_k}}, \tag{5}$$

for the probability of obtaining the collection in the order $\underline{\boldsymbol{i}}$ and the expected time to complete it in this order, respectively. We obtain bounds on $E[T(\underline{\boldsymbol{p}})]$ by computing $E(\cdot)$ for two extreme cases: the largest, when items are sampled in the most likely order (of decreasing probabilities), and in the reverse order, that produces the shortest possible expected completion time. The mean must be between these two, but the gap may be substantial and we are hard put to produce from it a tight estimate (in particular, the gap increases with m).

A recursive scheme of similar complexity is shown in [13]: let $E_{\{i_1, i_2, \cdots, i_k\}}$ denote the expected time to observe the entire sub-collection $\{i_1, i_2, \cdots, i_k\}$, then

$$E[T(\underline{\boldsymbol{p}})] = E_{N_m}, \tag{6}$$

and one proceeds recursively, from $E_{\{\varnothing\}} = 0$:

$$E_{\{i_1, i_2, \cdots, i_k\}} = \frac{1}{p_{i_1} + \cdots + p_{i_k}} + \sum_{r=1}^{k} \frac{p_{i_r}}{p_{i_1} + \cdots + p_{i_k}} E_{\{i_1, i_2, \cdots, i_k\} - \{i_r\}}. \tag{7}$$

Brayton obtains in [11] the expected time to complete k collections (equation (34) below), and the corresponding variance. He uses a different notation to express asymptotic properties: the $\{p_i\}$ are given by $\{F(i/n) - F((i-1)/n)\}$, and the distribution $F(\cdot)$ is assumed to have a nonvanishing density on [0, 1], with finite variation, and to achieve its minimal value at a finite number of isolated points. Naturally, the asymptotics turn out to hinge on this "minimum set."

Caron *et al.* [13] report on an effort to show the intuitive idea that $E[T(\underline{\boldsymbol{p}})]$ is minimized, over all probability vectors, in the EL case. This conjecture has been long part of the folklore of the CCP [1]. The result is interesting, since it gives a natural (and easily computable) yardstick by which to judge the relative difficulty of such problems. It is quite easy to show that $\underline{\boldsymbol{e}}/m$ is a stationary point of $E[T(\underline{\boldsymbol{p}})]$; the authors there also show—through equation (1)—that $\underline{\boldsymbol{e}}/m$ is a strong local minimum. But the proof that it is a global minimum remained elusive. The difficulty was that the common expressions for $E[T(\underline{\boldsymbol{p}})]$ are not convex in the components of $\underline{\boldsymbol{p}}$. While it appears that $E[T(\underline{\boldsymbol{p}})]$ *is* convex on the sheet $\Sigma p_i = 1$, there is no simple way to show this, uniformly for all m (in [13] this is shown for $m \le 6$ only).

Meilijson *et al.* develop in [36] expressions for the first two moments of the number of k-hits (defined below) and the number of drawings involved in them (these are the number of matches and of matched people, in the 'birthday problem'). They only consider the EL case and develop their formulas directly.

The articles [17] and [40], and especially the survey [12], consider statistical questions concerning the coupon collecting process, which we treat in Section 5.

The most detailed treatment of CCP-related questions is [44]. The authors' starting point is not any specific problem, but rather the solution: they consider a few parametrized families of Dirichlet integrals of type 2. They then show that the solutions of a very rich collection of sampling stopping-time problems can be expressed in terms of these integrals; in particular, that several CCP-related quantities fall squarely in that domain. [44] also contains numerous tables, and recurrence relations that can reduce higher-dimensional problems to the range of the tables. All the numerical data are geared to problems in which the p_i have at most two different values. In order to treat non-uniform sampling probability vectors, the authors provide Taylor expansions of the basic integrals at the EL point. Using those does not appear easy. The approach of the next section indicates that it should be possible to convert the above integrals directly to one-dimensional ones (possibly a sum of such integrals) for any probability vector.

Further comments on previous work appear below, when we discuss specific issues.

Computational Difficulty

We have commented on the huge computational effort required to obtain the expected value $E[T(\underline{\boldsymbol{p}})]$. It appears that except for the EL case, only rarely—and then, for rather small m—any expectations or probabilities have been explicitly calculated. We show below that asymptotics can sometimes be used, when $\underline{\boldsymbol{p}}$ is characterized by one or two parameters, but in general the components of $\underline{\boldsymbol{p}}$ do not satisfy any convenient relation and no useful asymptotic results exist (see the discussion of [10] in Section 4). The situation seemed so bad that a researcher in an area that applies the CCP complained that once m exceeds 30 or so, it is immaterial whether one knows the probabilities or not – since nothing can be computed with them anyway. In the next section we show that this is definitely not the case: we can routinely compute expectations and probabilities for thousands of items and more; the effort (for most of the computations) is roughly linear in m, with reasonable constants, independently of the probability values.

3. GENERATING FUNCTIONS FOR LANGUAGES

Flajolet *et al.* [21] present a method to compute probabilities for variables defined over sequences of independent samples from a finite population. They show it leads to a computationally feasible expression for the expected duration of the coupon collector activity. The same approach was briefly mentioned earlier by Comtet [15].

We outline of the method, and compute a few CCP-related quantities. In this section we provide analytic expressions, and take up computational considerations in Section 4.

Strings Over a Finite Alphabet

Let $A = \{a_1, \cdots, a_m\}$ be an *alphabet*, and each letter a_j is associated with the probability p_j. A string of letters from A is a *word.* The set of all finite words over A is denoted by A^*. A subset of A^* is called a *language*. Each word w has two types of weights: an additive weight − its size (the number of its letters), $|w|$, and a multiplicative weight − its probability, defined by: $\pi(w) \equiv \Pi_{j=1}^{|w|} p_{w_j}$. We define for a language L the following probability generating functions (PGF):

$$\phi_L(z) \equiv \sum_{w \in L} \pi(w) z^{|w|}, \qquad \hat{\phi}_L(z) \equiv \sum_{w \in L} \pi(w) \frac{z^{|w|}}{|w|!}. \tag{8}$$

The functions $\phi_L(z)$ and $\hat{\phi}_L(z)$ are called the ordinary and exponential PGFs of L, respectively. They are related through the Laplace-Borel transform:

$$\phi_L(z) = \int_{t \ge 0} \hat{\phi}_L(zt) e^{-t} dt. \tag{9}$$

Remark: These are not the standard PGFs of (discrete) random variables. Let $[x^r]f(x)$ denote the coefficient of x^r in the power series development of $f(x)$, then $[z^n]\ \phi_L(z)$ is the probability that a random word of size n is in the language L. This formalism does not support directly the notion of a random word of arbitrary size.

We define three operations on languages: union, catenation and shuffling.

1. The *union* of languages is the usual union of sets. We call it well-defined when the sets are *disjoint.*

2. The *catenation* of two languages, L_1 and L_2, is a language L, written as $L = L_1 L_2$, such that each word of L is formed by catenating a word from L_2 to a word from L_1, to form $w = w_1 . w_2$ (. is the catenation operator, that we often omit). The operation is well-defined iff L has the property of unique factorization.

Proposition 1: If the operations $L_U = L_1 \bigcup L_2$ and $L_C = L_1 L_2$ are well-defined, then

$$\phi_{L_U}(z) = \phi_{L_1}(z) + \phi_{L_2}(z), \qquad \phi_{L_C}(z) = \phi_{L_1}(z)\phi_{L_2}(z). \tag{10}$$

3. The operation of *shuffling* two languages is defined recursively as follows: Two languages are shuffled by shuffling all their words pair-wise; two words are shuffled by merging their letters in all possible manners, while retaining the original order in each ($\circ$ is the shuffle operator):

$$\begin{aligned} &w \circ \varepsilon = \varepsilon \circ w = w, && (\varepsilon \text{ is the null word.}) \\ &a.\, w_1 \circ b.\, w_2 = a.\,(w_1 \circ b.\, w_2) \bigcup b.\,(a.\, w_1 \circ w_2) && a, b \in A, \quad w_i \in L_i, \\ &L_1 \circ L_2 = \bigcup_{\substack{w_1 \in L_1 \\ w_2 \in L_2}} w_1 \circ w_2. && \end{aligned} \tag{11}$$

The operation is well-defined for languages that use disjoint subsets of A – that is, different alphabets.

Proposition 2: If the operation $L = L_1 \circ L_2$ is well-defined, then

$$\hat{\phi}_L(z) = \hat{\phi}_{L_1}(z)\hat{\phi}_{L_2}(z). \tag{12}$$

Applications of this tool have the following format: the statement of a problem is reinterpreted as a specification of a language $H \subseteq A^*$. The words of H are shown to be constructible from simple components by union, catenation and shuffling. The building-blocks are such that their PGFs are easy to compute directly, and propositions 1 and 2 provide the PGF of H, from which we 'read off' the desired answers. Here is a construction we use below.

Define a k-hit as the occurrence of a letter k times (or more) in a word. We compute the probability that a random word of a specified size has k-hits for exactly q distinct letters.

Let $H_{q,k}$ be the desired language: it only contains words in which exactly q letters recur at least k times, and the $m - q$ others appear at most $k - 1$ times. Introduce the following notation for one-letter languages:

$$a^{<k} = \{\varepsilon, a, a^2, \cdots, a^{k-1}\}, \qquad a^{\geq k} = a^k.\, a^*, \tag{13}$$

where a^2 is shorthand for the word aa, etc., and a^* is the supremum of all $a^{<k}$. Then

$$H_{q,k} = \bigcup_{I,J} \left(a_{i_1}^{\geq k} \circ a_{i_2}^{\geq k} \circ \cdots \circ a_{i_q}^{\geq k}\right) \circ \left(a_{j_1}^{<k} \circ a_{j_2}^{<k} \circ \cdots \circ a_{j_{m-q}}^{<k}\right). \tag{14}$$

The union is over all the partitions of A into two sets $I = \{i_1, \cdots, i_q\}$, $J = \{j_1, \cdots, j_{m-q}\}$, $I \bigcap J = \varnothing$, $I \bigcup J = N_m$. The exponential PGFs of the ingredients are straightforward. Let $e_k(z)$ denote the incomplete exponential function, $e_k(z) = \Sigma_{i=0}^{k} z^i/i!$, then the needed exponential PGFs are:

$$\hat{a}_i^{\geq k}(z) = \sum_{j \geq k} \frac{p_i^j z^j}{j!} = e^{zp_i} - e_{k-1}(zp_i), \qquad \hat{a}_i^{<k}(z) = e_{k-1}(zp_i). \tag{15}$$

The sets I and J involve disjoint alphabets; hence, the exponential PGF of $H_{q,k}$ is given by

$$\hat{\phi}_{q,k}(z) = \sum_{I,J} \prod_{i \in I} \left(e^{zp_i} - e_{k-1}(zp_i)\right) \prod_{j \in J} \left(e_{k-1}(zp_j)\right). \tag{16}$$

This sum has a more compact representation:

$$\hat{\phi}_{q,k}(z) = [u^q] \prod_{i=1}^{m} \left[e_{k-1}(zp_i) + u\left(e^{zp_i} - e_{k-1}(zp_i)\right)\right]. \tag{17}$$

If we are interested in probabilities of words of a specified size n, this is all that is needed. If the interest is in the sum of these probabilities over words of any size—or in CCP formulation: of any sequence of trials—we show below that we need the ordinary PGF, and the Laplace-Borel transform gives

$$\phi_{q,k}(z) = [u^q] \int_{t \geq 0} \prod_{i=1}^{m} \left[e_{k-1}(ztp_i) + u\left(e^{ztp_i} - e_{k-1}(ztp_i)\right)\right] e^{-t}\, dt. \tag{18}$$

In the EL case the summation in equation (16) is over $\binom{m}{q}$ identical terms, yielding

$$\hat{\phi}_{q,k}(z) = \binom{m}{q}\left(e^{z/m} - e_{k-1}(z/m)\right)^q \left(e_{k-1}(z/m)\right)^{m-q}, \qquad (\text{EL}) \tag{19}$$

and similarly for the ordinary PGF. We consider now a few CCP-related calculations.

(1) *The probability of drawing coupon #i at least r_i times in n trials,* $1 \le i \le m$

Each r_i is the *quota* of coupon #i. The number of drawings required to obtain the quotas $r_i = \delta n p_i$ for $\delta \in (0,1)$ is called the *δ-blanket time* for the process [47].

The samples satisfying this requirement form a language $H(\underline{\boldsymbol{p}}, \underline{\boldsymbol{r}})$, in which each word contains the letter a_i at least r_i times. Then in analogy with equation (14),

$$H(\underline{\boldsymbol{p}}, \underline{\boldsymbol{r}}) = \left(a_1^{\ge r_1} \circ a_2^{\ge r_2} \circ \cdots \circ a_m^{\ge r_m}\right). \tag{20}$$

where the $a_i^{\ge r_i}$ are defined in equation (13). The exponential PGF of the set $a_i^{\ge r_i}$ is $e^{p_i z} - e_{r_i - 1}(p_i z)$, with $e_{-1}(\cdot) \equiv 0$; hence

$$\hat{\phi}(\underline{\boldsymbol{p}}, \underline{\boldsymbol{r}}; z) = \prod_{i=1}^{m} \left(e^{p_i z} - e_{r_i - 1}(p_i z)\right), \tag{21}$$

and the desired probability is given by

$$\mathrm{P}(\underline{\boldsymbol{p}}, \underline{\boldsymbol{r}}; n) = n! [z^n] \prod_{i=1}^{m} \left(e^{p_i z} - e_{r_i - 1}(p_i z)\right). \tag{22}$$

There is no useful explicit form for this coefficient. For moderate m and n, its numerical value may be obtained using the Cauchy integral formula with any contour around the origin, since $\hat{\phi}(\underline{\boldsymbol{p}}, \underline{\boldsymbol{r}}; z)$ is an entire function.

An equivalent form is obtained in [42, p.51]. A relatively efficient numerical recursion to evaluate $\mathrm{P}(\underline{\boldsymbol{p}}, \underline{\boldsymbol{r}}; n)$ is developed by Neuts and Carson [37].

(2) *The probability of scoring k-hits for r coupons in a sample of size n*

The construction is similar to the one above; since the items are not specified, we need to sum over all possible r-out-of-m sets.

What is the probability of scoring *at least r k*-hits? To use equation (17), define $Y_n(\underline{\boldsymbol{p}}, k)$ as the number of k-hits in a word of size n. Then

$$\begin{aligned} \Pr[Y_n(\underline{\boldsymbol{p}}, k) = j] &= n! [z^n u^j] \prod_{i=1}^{m} \left(e_{k-1}(p_i z) + u(e^{p_i z} - e_{k-1}(p_i z))\right) \\ &= [z^n u^j] \int\limits_{t \ge 0} \prod_{i=1}^{m} \left(e_{k-1}(p_i z t) + u(e^{p_i z t} - e_{k-1}(p_i z t))\right) e^{-t} dt. \end{aligned} \tag{23}$$

The required answer is $\Sigma_{r \le j \le m} \Pr[Y_n(\underline{\boldsymbol{p}}, k) = j]$. Unless r is a very small number – or similarly close to m, there is no essentially simpler way to express this truly complex combinatorial quantity. The relative simplicity of the EL case is probably more apparent in this case than in others (see [36]).

(3) *The expected number of different coupons drawn in a sample of size n*

This problem calls for the evaluation of $E[Y_n(\underline{\boldsymbol{p}}, 1)]$. It can be done by elementary considerations, but it is instructive to use the present apparatus to recapture it. From equation (23) we have

$$E[Y_n(\underline{\boldsymbol{p}}, 1)] = \sum_{j=1}^{n} j n! [z^n u^j] \prod_{i=1}^{m} \left(1 + u(e^{p_i z} - 1)\right) \equiv n! [z^n] \frac{\partial}{\partial u} b(u, z)\Big|_{u=1}, \tag{24}$$

where $b(u,z) = \Pi_{i=1}^{m} \left(1 + u(e^{p_i z} - 1)\right)$. The differentiation and evaluation are routine:

$$\begin{aligned}\frac{\partial}{\partial u}\, b(u,z)\Big|_{u=1} &= b(u,z)\sum_{i=1}^{m}\frac{e^{p_i z}-1}{1+u(e^{p_i z}-1)}\Big|_{u=1} \\ &= b(1,z)\sum_{i=1}^{m}\frac{e^{p_i z}-1}{e^{p_i z}} \;=\; e^{z}\sum_{i=1}^{m}\Big(1-e^{-p_i z}\Big).\end{aligned} \tag{25}$$

Leading to the obvious value

$$E[Y_n(\underline{\boldsymbol{p}},1)] = n!\sum_{i=1}^{m}\Big(\frac{1}{n!}-\frac{(1-p_i)^n}{n!}\Big) = \sum_{i=1}^{m}[1-(1-p_i)^n], \tag{26}$$

We consider some properties of this result below and in Section 4.

(4) *The expected time to draw a sub-collection of size j* [21]

Consider the following set of equalities:

$$E[T_j(\underline{\boldsymbol{p}})] = \sum_{n\ge 0}\Pr[T_j(\underline{\boldsymbol{p}})>n] = \sum_{n\ge 0}\Pr[Y_n(\underline{\boldsymbol{p}},1)<j], \tag{27}$$

which hold since the two compound events $\{T_j(\underline{\boldsymbol{p}})>n\}$ and $\{Y_n(\underline{\boldsymbol{p}},1)<j\}$ consist of the same sequences of trials. To use equation (23) we write

$$\sum_{n\ge 0}\Pr[Y_n(\underline{\boldsymbol{p}},1)<j] = \sum_{n\ge 0}\sum_{r=0}^{j-1}\Pr[Y_n(\underline{\boldsymbol{p}},1)=r] = \sum_{r=0}^{j-1}\{\,\sum_{n\ge 0}\Pr[Y_n(\underline{\boldsymbol{p}},1)=r]\}. \tag{28}$$

Equation (23) now gives, specialized to $k=1$,

$$E[T_j(\underline{\boldsymbol{p}})] = \sum_{r=0}^{j-1}\sum_{n\ge 0}\Pr[Y_n(\underline{\boldsymbol{p}},1)=r] = \sum_{r=0}^{j-1}[u^r]\int_{t\ge 0}\prod_{i=1}^{m}\Big(1+u(e^{p_i t}-1)\Big)e^{-t}dt. \tag{29}$$

The integrand is an m-degree polynomial in u, and when $j=m$ it simplifies greatly: in this case we need the sum of the coefficients of u^r for all r except $r=m$, which is $\Pi_{i=1}^{m}(e^{p_i t}-1)$. The sum of all the coefficients is simply the value of the right-hand side at $u=1$, and we find:

$$E[T_m(\underline{\boldsymbol{p}})] = E[T(\underline{\boldsymbol{p}})] = \int_{t\ge 0}\Big[e^t-\prod_{i=1}^{m}(e^{p_i t}-1)\Big]e^{-t}dt = \int_{t\ge 0}\Big[1-\prod_{i=1}^{m}\Big(1-e^{-p_i t}\Big)\Big]dt. \tag{30}$$

This is a computationally efficient form for $E[T(\underline{\boldsymbol{p}})]$. The equality of this integral to the right-hand side of equation (1) is immediate. Equation (30) has a particularly simple form in the EL case,

$$E[T(\underline{\boldsymbol{e}}/m)] = \int_{t\ge 0}\Big[1-\Big(1-e^{-t/m}\Big)^m\Big]dt. \tag{31}$$

There is interest also in finding the time to ‘nearly complete’ the sampling, that is, for values of j that are very close to m, and the same approach applies. For example, we find

$$E[T_{m-1}(\underline{\boldsymbol{p}})] = \int_{t\ge 0}\Big[1-\sum_{j=1}^{m}\Big(p_j+e^{-p_j t}(1-p_j)\Big)\prod_{\substack{i=1\\ i\ne j}}^{m}\Big(1-e^{-p_i t}\Big)\Big]dt. \tag{32}$$

With some care the computation times of these quantities can be still kept essentially linear in m.

Remark: It is instructive to consider side by side the two quantities given by equations (26) and (29). Both may be viewed as functions in the (sampling duration, number of captures) plane, but with the relation of independent/dependent variables reversed, as shown for a special case in Fig. 1.

The engineering (and statistical) significance of the two functions is very different: The “detection curve,” given by equation (26), shows the expected fraction of detected items as a function of the sampling duration. The “sample duration curve,” computed in equation (29), gives the expected duration required to complete a specified fraction. Equations (30) and (32) above give expressions for two points on this curve; computing each is linear in m, but obtaining the entire curve still appears infeasible when the number of distinct probabilities is large. Note that the first curve extends to infinity (along the sample size axis), while the second one terminates in the point $(E[T(\underline{\boldsymbol{p}})], 1)$. Statistical inference is joined to the detection curve in Section 5, where we discuss the prediction function. Another way to distinguish the two curves is to consider how they compare with individual experiments. Points of the first curve represent average of samples scattered along the ordinate, while the second one features averages of horizontal scatter.

We *conjecture* that the appearance of Fig. 1, where the duration curve lies entirely above the detection curve (except for the first two points: $t = 0, 1$, where they coincide) is unique to the EL case, and that in all other cases they intersect.

The next two problems generalize problem (4) in different ways.

(5) *The expected time to draw k times a sub-collection of size j*

When $k = 1$ this repeats the previous problem. For $j = m$, $k = 2$ and the EL case, this problem is known as the "double Dixie cup problem" (after the product that carried the collected coupons). Holst provides an answer for the EL case in [31], and comments on earlier derivations of that result, extending to 1960, as a rule considering complete collections (i.e. $j = m$), and—except [11]—the EL case. Other references of interest (all essentially concerned with limiting properties of $E[T(\underline{\boldsymbol{e}}/m)]$) are [23, 33, 38 and 39].

The required expected value is obtained exactly as for Problem (4):

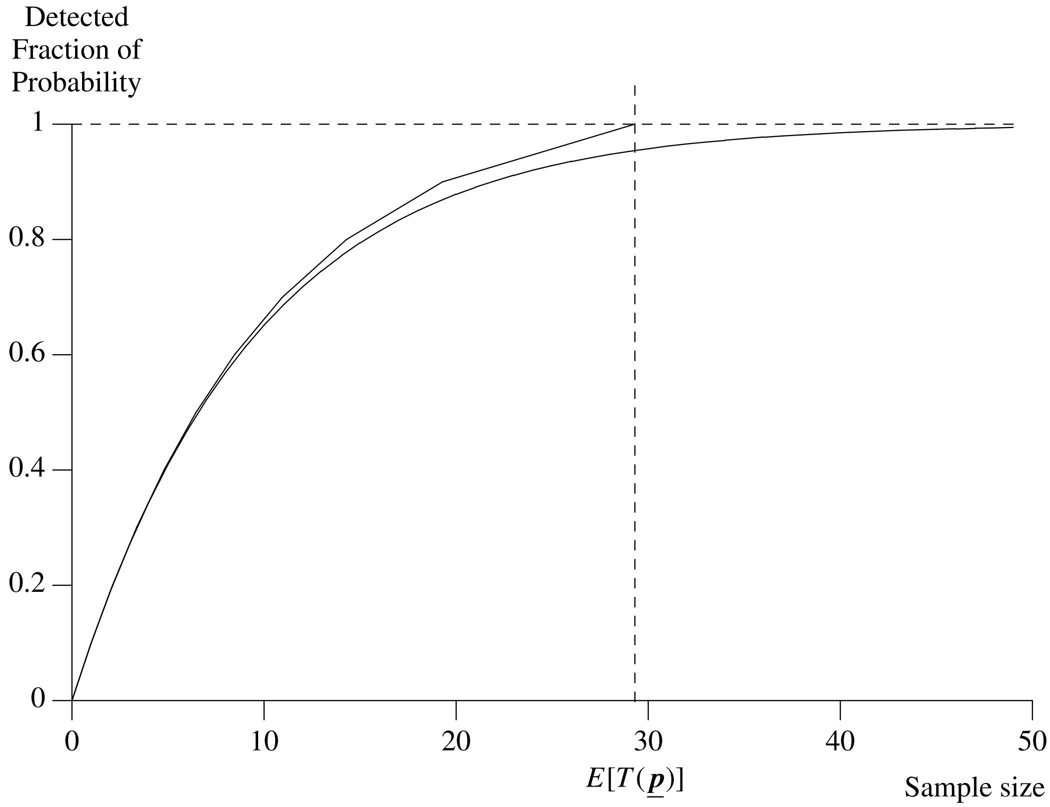

Fig. 1: Detection−temporal relationships for the EL case (m = 10)

$$E[T_j^{(k)}(\underline{\boldsymbol{p}})] = \sum_{n\ge 0} \Pr[T_j^{(k)}(\underline{\boldsymbol{p}}) > n] = \sum_{n\ge 0} \Pr[Y_n(\underline{\boldsymbol{p}}, k) < j]. \tag{33}$$

The same approach that produced equations (29) and (30) gives here

$$E[T_j^{(k)}(\underline{\boldsymbol{p}})] = \sum_{r=0}^{j-1} \sum_{n\ge 0} \Pr[Y_n(\underline{\boldsymbol{p}}, k) = r] = \sum_{r=0}^{j-1} [u^r] \int_{t\ge 0} \prod_{i=1}^{m} \Big(1 + u(e^{p_i t} - e_{k-1}(p_i t))\Big) e^{-t}\, dt,$$

and

$$E[T^{(k)}(\underline{\boldsymbol{p}})] = \int_{t\ge 0} \Big[1 - \prod_{i=1}^{m} \Big(1 - e^{-p_i t} e_{k-1}(p_i t)\Big)\Big] dt. \tag{34}$$

Computing this is roughly of the same difficulty as that of $E[T(\underline{\boldsymbol{p}})]$, and increases sub-linearly with k. Equation (34) was obtained in [11] by direct appeal to the multinomial distribution of the drawings.

(6) *The expected time to draw a specified sub-collection C of size $|C| = j$*

This problem can be handled almost as problem (4), but is much more informative, as we show later. Let the required subset be $C \subseteq A$. As above,

$$E[T_C(\underline{\boldsymbol{p}})] = \sum_{n\ge 0} \Pr[T_C(\underline{\boldsymbol{p}}) > n] = \sum_{n\ge 0} \Pr[W_n(C, \underline{\boldsymbol{p}}) < j] = \sum_{k=0}^{j-1} \{ \sum_{n\ge 0} \Pr[W_n(C, \underline{\boldsymbol{p}}) = k]\}, \tag{35}$$

where $W_n(C, \underline{\boldsymbol{p}})$ is the number of coupons from C observed in a sequence of n drawings. Let $H_k(C)$ be a language containing only words in which exactly k items of C appear. Using equations (13) and (14) it can be written as

$$H_k(C) = \bigcup_{I \subseteq C} \left(a_{i_1}^{\ge 1} \circ a_{i_2}^{\ge 1} \circ \cdots \circ a_{i_k}^{\ge 1}\right) \circ (A - C)^*. \tag{36}$$

The union is over all $I = \{a_{i_1}, \cdots, a_{i_k}\} \subseteq C$; hence $H_k(C)$ has the exponential PGF

$$\hat{\phi}_{k,C}(z) = \sum_{I:\, |I|=k} \prod_{i\in I} \left(e^{zp_i} - 1\right) e^{z(1-P_C)}, \tag{37}$$

where $P_C = \sum_{i\in C} p_i$. Let us write $C = \{a_{(1)}, \cdots, a_{(j)}\}$. The same manipulations that led us to equation (30) provide

$$\hat{\phi}_{k,C}(z) = [u^k] \prod_{i=1}^{j} \Big[1 + u\Big(e^{zp_{(i)}} - 1\Big)\Big] e^{z(1-P_C)}, \tag{38}$$

and for the expectation we find

$$\begin{aligned} E[T_C(\underline{\boldsymbol{p}})] &= \sum_{k=0}^{j-1} \phi_{k,C}(1) = \sum_{k=0}^{j-1} \int_{t\ge 0} [u^k] \prod_{i=1}^{j} \Big[1 + u\Big(e^{tp_{(i)}} - 1\Big)\Big] e^{t(1-P_C)} e^{-t}\, dt, \\ &= \int_{t\ge 0} \Big[1 - \prod_{(i)\in C} \Big(1 - e^{-p_{(i)}t}\Big)\Big] dt, \end{aligned} \tag{39}$$

in complete analogy with equation (30).

(7) *Tail probabilities for $T(\underline{\boldsymbol{p}})$*

We have obtained expressions for the distribution of $T(\underline{\boldsymbol{p}})$: from the discussion leading to $E[T(\underline{\boldsymbol{p}})]$ – see equation (27) – we find for the tail probabilities of $T(\underline{\boldsymbol{p}})$, specializing equation (23) to $k = 1$, that

$$\Pr[T(\underline{\boldsymbol{p}}) > n] = \Pr[Y_n(\underline{\boldsymbol{p}}, 1) < m] = \sum_{r=0}^{m-1} n! [z^n u^r] \prod_{i=1}^{m} \Big(1 + u(e^{p_i z} - 1)\Big). \tag{40}$$

The same treatment that led to equation (30) then provides

$$\Pr[T(\underline{\boldsymbol{p}}) > n] = n![z^n]\Big(e^z - \prod_{i=1}^{m}(e^{p_i z} - 1)\Big) = 1 - n![z^n]\prod_{i=1}^{m}(e^{p_i z} - 1). \tag{41}$$

This is a specialization of equation (22), and we discuss it further in Section 4.

Consider now another descriptor of the distribution – the variance $V[T(\underline{\boldsymbol{p}})]$. We use the relation $E[T^2(\underline{\boldsymbol{p}})] = \Sigma_{n\ge 0}(2n+1)\Pr[T(\underline{\boldsymbol{p}}) > n] = 2I_2 + E[T(\underline{\boldsymbol{p}})]$. To compute I_2 we have the choice of using the exponential PGF or the ordinary one. The first produces

$$I_2 = \sum_{n\ge 1} n\Pr[T(\underline{\boldsymbol{p}}) > n] = \sum_{n\ge 1} n\Big(1 - n![z^n]\prod_{i=1}^{m}(e^{p_i z} - 1)\big)\Big). \tag{42}$$

The second one yields

$$\begin{aligned} I_2 = \sum_{n\ge 1} n\Pr[T(\underline{\boldsymbol{p}}) > n] &= \int\limits_{t\ge 0} \sum_{j=1}^{m} p_j t\Big[1 - \prod_{\substack{i=1\\ i\ne j}}^{m}\big(1 - e^{-p_i t}\big)\Big]dt. \\ &= \int\limits_{t\ge 0} \sum_{j=1}^{m} p_j t\Big[1 - \frac{\Pi(t)}{1 - e^{-p_j t}}\Big]dt, \qquad \Pi(t) \equiv \prod_{i=1}^{m}\big(1 - e^{-p_i t}\big). \end{aligned} \tag{43}$$

Then $V[T(\underline{\boldsymbol{p}})] = 2I_2 + E[T(\underline{\boldsymbol{p}})](1 - E[T(\underline{\boldsymbol{p}})])$. The first expression for I_2 is similar to equation (1), and there is little to recommend it. The second is numerically feasible, very similar to the integral used for $E[T(\underline{\boldsymbol{p}})]$; with some care the computation time can be kept linear in m, although with higher constants.

(8) *On the distribution of* $N_1(n)$

$N_1(n)$ is the number of letters that occur exactly once in a word of size n; it is used in the statistical considerations below. The words that contribute to $\Pr[N_1(n) = r]$ are similar to those we constructed in problem (3), and we find the language and generating function

$$H(\underline{\boldsymbol{p}}, r) = \bigcup_{I:\,|I|=r} \Big[\big(a_{i_1} \circ a_{i_2} \circ \cdots \circ a_{i_r}\big) \circ (A - I)^{\ne 1}\Big], \tag{44}$$

$$\hat{\phi}_r(z) = \sum_{I:\,|I|=r} \prod_{i\in I} z p_i \prod_{j\notin I}\big(e^{z p_j} - z p_j\big). \tag{45}$$

Hence

$$\Pr[N_1(\underline{\boldsymbol{p}}, n) = r] \; = \; n![z^n u^r]\prod_{i=1}^{m}\big(e^{z p_i} + (u-1) z p_i\big). \tag{46}$$

In general, numerical evaluations are not simple, and unless the components of $\underline{\boldsymbol{p}}$ have an analytically tractable form asymptotic estimates are unlikely. For the special case $r = n$, a recursion is shown in [22], as well as a few ad-hoc approximations.

The expectation is straightforward either from equation (46) or elementary considerations, and equals

$$E[N_1(\underline{\boldsymbol{p}}, n)] \; = \; n\sum_{i=1}^{m} p_i(1 - p_i)^{n-1}. \tag{47}$$

A similar procedure leads to

$$E[N_k(\underline{\boldsymbol{p}}, n)] \; = \; \binom{n}{k}\sum_{i=1}^{m} p_i^k(1 - p_i)^{n-k}. \tag{48}$$

(9) *Proof of the Conjecture on the extremality of the* EL *case*

The conjecture that $E[T(\underline{\boldsymbol{p}})]$ is minimal when the probability vector $\underline{\boldsymbol{p}}$ is uniform is of interest, because it provides an easily computable lower bound when the actual probabilities are unknown. We start with equation (30) and denote the integrand by $f(t, \underline{\boldsymbol{p}})$. This function is everywhere positive, and we show that it is minimized—uniformly in t—at the probability vector $\underline{\boldsymbol{e}}/m$. The desired result follows. The minimization problem

$$\min_{\underline{\boldsymbol{p}}\geq 0} f(t, \underline{\boldsymbol{p}}) = \min_{\underline{\boldsymbol{p}}\geq 0}\Big[1 - \prod_{i=1}^{m}\Big(1 - e^{-p_i t}\Big)\Big] \qquad \text{Subject to } \sum_{i=1}^{m} p_i = 1, \tag{49}$$

is equivalent to the problem

$$\max_{\underline{\boldsymbol{p}}\geq 0}[1 - f(t, \underline{\boldsymbol{p}})] = \max_{\underline{\boldsymbol{p}}\geq 0} \prod_{i=1}^{m}\Big(1 - e^{-p_i t}\Big) \qquad \text{Subject to } \sum_{i=1}^{m} p_i = 1. \tag{50}$$

The optimum is at some strictly positive $\underline{\boldsymbol{p}} > 0$. Then $f(t, \underline{\boldsymbol{p}})$ is positive, and since the logarithm function is monotone increasing over the positive reals, the above problem is equivalent to

$$\max_{\underline{\boldsymbol{p}}>0} \ \ln[1 - f(t, \underline{\boldsymbol{p}})] = \max_{\underline{\boldsymbol{p}}>0} \sum_{i=1}^{m} \ln\Big(1 - e^{-p_i t}\Big) \qquad \text{Subject to } \sum_{i=1}^{m} p_i = 1. \tag{51}$$

The last problem is equivalent to searching for the saddle-point of the Lagrangian

$$L(p, \lambda, \underline{\boldsymbol{v}}) \ \equiv \ \sum_{i=1}^{m} \ln\Big(1 - e^{-p_i t}\Big) + \lambda\Big(\sum_{i=1}^{m} p_i - 1\Big) + \sum_{i=1}^{m} v_i p_i. \tag{52}$$

For any $t > 0$ the Lagrangian $L(p, \lambda, \underline{\boldsymbol{v}})$ has a stationary point at the uniform $\underline{\boldsymbol{p}}$ (with the values for λ and $\underline{\boldsymbol{v}}$ there uniquely defined). Moreover: it is everywhere concave in $\underline{\boldsymbol{p}}$-space (its Hessian is diagonal, with negative elements only), hence that point is a global maximum for equation (50), and a global minimum for $E[T(\underline{\boldsymbol{p}})]$.

S.M. Ross has pointed out [41] that a stronger claim can be shown: $T(\underline{\boldsymbol{p}})$ is minimized *stochastically*[1] when $\underline{\boldsymbol{p}} = \underline{\boldsymbol{e}}/m$. We compute $\Pr[T \leq n]$ when $\underline{\boldsymbol{p}}$ has at least 2 unequal components, say $p_1 \neq p_2$. Let $\underline{\boldsymbol{p}}' = (q, q, p_3, \cdots, p_m)$, where $q = (p_1 + p_2)/2$. Then

$$\Pr[T \leq n] = \sum_{\underline{\boldsymbol{r}}_n} \Pr[T \leq n \mid \text{coupon } \#i \text{ is selected } r_i \text{ times}, \ \ 3 \leq i \leq m] \times \Pr[\underline{\boldsymbol{r}}_n],$$

where $\underline{\boldsymbol{r}}_n$ is the number of times the specified coupons are obtained during the first n trials. The unconditional probability is the same for $\underline{\boldsymbol{p}}$ and $\underline{\boldsymbol{p}}'$, and the conditional one vanishes unless all $r_i > 0$ and $s \equiv n - \Sigma\, r_i$ satisfies $s \geq 2$. In that case it equals $1 - [p_1/(p_1 + p_2)]^s - [p_2/(p_1 + p_2)]^s$ with $\underline{\boldsymbol{p}}$, and $1 - 1/2^{s-1}$ with $\underline{\boldsymbol{p}}'$. The convexity of the function x^s implies that $\Pr[T \leq n]$ is larger with $\underline{\boldsymbol{p}}'$, and the argument proceeds similarly to "equalize" the rest of the components. This is equivalent to showing that $\Pr[T \leq n]$ is a Schur-convex function of $\underline{\boldsymbol{p}}$ [34]. □

Many similar questions are of operational interest and can be answered by these means. Sobel *et al.* [44] bring up a few more questions that they tackle by their approach. Its computational feasibility depends on special circumstances, such as $\underline{\boldsymbol{p}}$ having few different components. Here is an example that gives the flavor of those questions:

> Given two sets of items, C and D, what is the probability of finding r of the first before capturing k of the second?

Improvements of the algorithm PREDUCE described in Section 1 lead to CCPs with the multiple completion-sets criterion: Let $\{A_i\}$, $1 \leq i \leq r$ be r disjoint subsets of A. What is the number of trials required for the elements of at least one of them to be all obtained? For such problems the preferred route is via the Poisson transform of Section 5.

[1] A *rv* X is said to be stochastically smaller than Y when $\Pr[X \leq x] \geq \Pr[Y \leq x]$ for all x. This is denoted by $X \leq_{st} Y$.

4. COMPUTATIONAL ASPECTS AND NUMERICAL EXAMPLES

A major consideration in the treatment presented above is its suitability for numerical work. We describe some computations, and remark on the significance of the numerical results. For engineering problems modeled by the CCP the following quantities are commonly needed:

(a) The expectation of $T(\underline{\boldsymbol{p}})$.

(b) Dispersion measures of $T(\underline{\boldsymbol{p}})$, where tail probabilities are possibly the most useful.

(c) The tradeoffs between length of sampling and the probabilities of detecting given fractions of the items.

(d) Practical considerations in applications where the vector $\underline{\boldsymbol{p}}$ is *unknown*.

(a) *The expectation of* $T(\underline{\boldsymbol{p}})$

The computational advantage of the integral in equation (30) over the sums of Section 2 is enormous: instead of dealing with 2^m oscillating terms, we integrate a bounded, smooth, everywhere-positive function. It is true that the range of integration may be tremendous as well, but the function is smooth enough for an integration routine with locally adaptive step-size to compute it with a few hundreds of function evaluations, under very stringent accuracy requirements.

This convenience was already noted in [21] and was the starting point of our interest in the issue, as it computes an ε-approximation in linear time (in m), for a problem that is inherently exponential in m. Fig. 2 shows the integrand $f(t)$ of equation (30) for the EL case with $m = 100$, and the shape is typical.

The curve starts from 1 at $t = 0$ and decreases very slowly (its first $m-1$ derivatives vanish at $t = 0$), has a relatively restricted region of faster descent towards 0, where it has an inflection point, and vanishes exponentially. The dominant term in the tail is $\exp(-p_{\min}t)$, where $p_{\min}$ is the smallest element in $\underline{\boldsymbol{p}}$. This fact was used to determine the cut-off point for the integration, t_u. Requiring an absolute error on the order of ε, t_u was determined by the relation

$$\int_{t\ge t_u} e^{-tp_{\min}}dt = \frac{1}{p_{\min}}\, e^{-t_u p_{\min}} = \varepsilon, \tag{53}$$

or $t_u = -\ln(p_{\min}\varepsilon)/p_{\min}$. If there are r items associated with the minimal probability $p_{\min}$, the cut-off point needs to be only slightly pushed to $t_u = -\ln(p_{\min}\varepsilon/r)/p_{\min}$. We can replace the infinite range of integration by a finite one, [0,1], with the substitution $t = -\ln x$, and get a singular integrand for our pains. For practical numerical work we are probably better off this way.

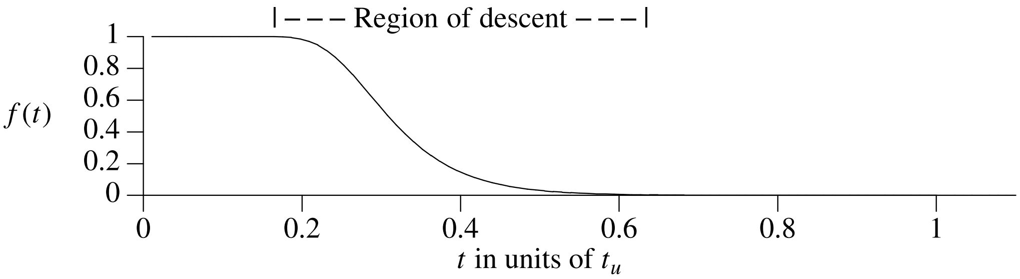

Fig. 2: The integrand of equation (30)

We routinely integrated along the interval $(t_u, 1.5t_u)$, to estimate the actual error. The value of ε was set at 10^{-7}, far smaller than is necessary in applications, and was selected to strain the integration procedures. The integration was done with the routine *quanc8* of [24], which uses 8-point Newton-Cotes integration with locally-adaptive step size. Sample results are shown in tables 1 and 2.

m	10	100	1000	10,000	100,000
$E[T(\underline{p})]$	56.24	1856.52	41,288.83	739,709.57	11,670,512.01
$m(\ln m)^2$	53.02	2120.76	47,717.1	848,303.7	13,254,745
t_u	571	17,406	281,140	4,053,471	54,629,467
# Evalu.	177	161	177	193	145

Table 1: $E[T(\underline{p})]$ for the Zipf distribution.

The probability vector used in table 1 is the Zipf distribution, with $p_i = 1/iH_m$. The second line corresponds to the estimate $m\ln^2 m$, shown in [10] to be an asymptotic estimate of $E[T(\underline{p}_{Zipf})]$. The estimate clearly tracks the correct values, with an overestimate that decays extremely slowly. The computational effort is roughly linear in m (and is independent of the particular probability vector, so long as all the probabilities are distinct). The last line reports the number of function evaluations used in the integration. The oscillations in this number depend on the location of the region of descent of the integrand with respect to the evaluation points selected by the integration routine; some fine-tuning is possible here but was not deemed worthwhile.

m	10	100	1000	10,000	100,000
$E[T(\underline{p})]$	68.985	6338.74	628,226.33	62,766,148.84	6,276,050,044.96
$E[T(\underline{e}/m)]$	29.929	518.74	7485.47	97,876.06	120,901.46
t_u	1107	124,458	1.464×10^7	1.692×10^9	1.923×10^{11}
# Evalu.	177	225	225	209	209

Table 2: $E[T(\underline{p})]$ for the Linear distribution.

Table 2 presents results for a different distribution that arises in applications: the so-called Linear distribution. There $p_i = 2i/m(m+1)$. For comparison, the second line brings the corresponding expected values in the EL case. This distribution is the reverse, in a way, of the Zipf distribution: there as here we have many probabilities of very close values, but for the Zipf distribution it is the small probabilities that are close, whereas for the linear one it is the higher values that are nearly uniform, and at the tail they are well spaced out. The change this causes in $E[T(\underline{p})]$ is dramatic: since it must be in $\Omega(1/p_{\min})$, it is at least quadratic in m. In this case we can produce better asymptotics, at $m(m+1)(2\pi/\surd 3-3) \approx 0.6275987m(m+1)$. This result was stated first in [16]. The proof consists in showing that with this $\underline{p}$ the integral in equation (30) can be written in the limit $m\to\infty$ as $\int_0^1 \Sigma_{k\geq 1} b_k x^k dx$, where b_k is a well-known number-theoretic function, specifying the difference between the number of partitions of k with even number of distinct parts, and the number of such odd-sized partitions. In [27, p.14] it is shown that b_k is $(-1)^n$ when k equals $n(3n+1)/2$ for some integer n, and vanishes otherwise (essentially Euler's pentagonal numbers theorem). The integration is routine. The asymptotic estimate agrees with the exact value to six decimal places already for $m = 100$.

Remark: The influence of the last few hard-to-get items on the expected sampling time is considerable: To demonstrate the effect of those rare items, we used equation (39) with $m = 1000$ to compute the expected time with the Linear distribution to draw all but the 10 items with smallest probabilities, and obtained $E[T_{C_{990}}] = 106,387.30$, about one sixth of the corresponding $E[T(\underline{p})]$.

The smoothness of the function we integrate does not lend it to asymptotic analysis by most standard methods. Recently the issue was investigated by S. Boneh and Papanicolaou in [10]. Their main result is establishing two types of behavior for this value: Let $\underline{p}$ be the normalized m-prefix of the infinite sequence $\alpha = \{a_n\}$. They show that the limit $\lim_{m\to\infty} \int [1-\Pi_{n=1}^m \big(1-\exp(-a_n t)\big)]dt$ exists, and denote it by $L(\alpha)$. When $L(\alpha)$ is finite, $E[T(\underline{p})]$ is naturally $\sim L(\alpha)\Sigma_{n=1}^m a_n$. When $L(\alpha)=\infty$, they show that under additional restrictions on the a_n, the $L(\alpha)$ in the expression for $E[T(\underline{\boldsymbol{p}})]$ can be replaced by $f(m)\ln[f(m)/f'(m)]$, where $f(n) \equiv 1/a_n$. The linear, as well as the geometric distribution fall in the first type; there seems to be no general method to obtain an estimate for $L(\alpha)$ in this case. For example, all we can say about sampling with $p_i = cq^i$ (with $q > 1$) is that $E[T(\underline{\boldsymbol{p}}_{Geom})]$ is proportional to $q^{m+1}L(\{q^n\})$, which reminds us of equation (53).

(b) *Dispersion measures and tail probabilities*

The variance of $T(\underline{\boldsymbol{p}})$ is somewhat more expensive to calculate than the expected value – even when using equation (43). Hence we only computed it for $m = 1000$ for the above distributions, and evaluated the variance ratios. We found the values 0.20521 and 0.72201 for the Zipf and the Linear distributions, respectively. (For the EL case it equals $\sqrt{1637450}/7485 = 0.17134$). Tail probabilities are hard to compute, despite the innocent appearance of equation (41). A direct approach uses the Cauchy integral formula,

$$\Pr[T(\underline{\boldsymbol{p}}) > n] = 1 - n! \oint_C z^{-n-1} \prod_{i=1}^{m} (e^{p_i z} - 1)dz = 1 - \frac{n!}{2\pi} \int_0^{2\pi} z^{-n} \prod_{i=1}^{m} (e^{p_i z} - 1)d\theta, \tag{54}$$

where in the last expression $z = e^{i\theta}$. The integrand is very volatile, hence direct integration is unstable. However, the integrand is significant only at values of z with small argument, and we can use the saddle-point method: writing the integrand in the first line of equation (54) as $\exp(h_n(z))$, we have

$$\Pr[T(\underline{\boldsymbol{p}}) > n] \approx 1 - n! \frac{\prod_{i=1}^{m} (e^{p_i R_n} - 1)}{R_n^{n+1}\sqrt{2\pi h_n{}''(R_n)}}, \tag{55}$$

where R_n is the root of the equation $h_n'(z) = 0$. Experiments with this expression showed it is difficult to obtain meaningful results— such as tail probabilities below 0.02—without using (the relatively slow) multiprecision arithmetic. Consider for example the EL case, where it is easy to show that $\Delta \equiv n + 1 - R_n$ is positive, and tends fast to zero (it is smaller than 1 already for $n \approx m \ln m$, that is – for all values at which one would consider computing tail probabilities). We find there

$$\Pr[T(\underline{\boldsymbol{p}}) > n] \approx 1 - \frac{n! R_n^{m-n}}{\Delta^m \sqrt{2\pi(n+1)(1 - \Delta/m)}} \qquad (\text{EL}). \tag{56}$$

Solving for R_n is easy, but the cancellation is time-consuming to overcome.

(c) *Sampling tradeoffs*

Consider equation (26). We can get closed-form results from it for the EL case as follows. Picking $n = E[T(\underline{\boldsymbol{e}}/m)] = mH_m$ we find, for large m

$$E[Y_n(\underline{\boldsymbol{e}}/m, 1)] = m[1 - (1 - \frac{1}{m})^{mH_m}] \approx m(1 - e^{-H_m}). \tag{57}$$

The approximation of H_m by $\ln m + \gamma$ suffices here; the error is $(12m)^{-1} + O(m^{-2})$, while $\gamma \approx 0.57722$. We find that the expected number of items detected by the time the collector would expect to finish is *extremely* close to m, at $m - e^{-\gamma} \approx m - 0.56146$, with the shortfall essentially independent of m (Fig. 1 suggests this as well). From equation (2), the standard deviation of $T(\underline{\boldsymbol{e}}/m)$ is approximately $m\pi/\sqrt{6} \approx 1.28255m$. It is a suitable unit of comparison with $E[T(\underline{\boldsymbol{e}}/m)]$. In Table 3 we denote the expected number of items found in $E[T(\underline{\boldsymbol{e}}/m)] + k \times m\pi/\sqrt{6}$ trials by $m - a$, where a the ‘shortfall’.

k	a
–4	94.9148
–3	26.3227
–2	7.30004
–1	2.02451
0	0.56146
1	0.15571
2	0.04318
3	0.01198
4	0.00332

Table 3: Expected shortfalls for sampling in the EL case

To appreciate the values in this table, note first that these shortfalls are virtually independent of m. Secondly, the length of $E[T(\underline{\boldsymbol{e}}/m)]$, when measured in standard deviations is quite small: it comes to 3.6 for $m = 100$, 5.386 for $m = 1000$ (where all of the approximations used above are quite tight), and 7.1813 and 8.9766 for m at 10^4 and 10^5, respectively. For the last case, e.g., the table provides that in the first 55% of the expected total sampling time, the expected number of detected items is 99905.1. On the other hand, to be fairly confident that all the items are obtained the collector must endure a very long sampling sequence.

For other vectors $\underline{\boldsymbol{p}}$ we do not have such closed expressions, but numerical experimentation revealed very similar patterns, usually with smaller shortfalls. Formally this is so because of their larger variance ratios, but it is inutitively clear as well, since T there is almost entirely due to the smaller $p_i \ll 1/m$.

(d) *Unknown probability vector* $\underline{\boldsymbol{p}}$

We look now at the CCP when the vector $\underline{\boldsymbol{p}}$ is not known. To our mind, this is the most interesting situation, since this is the state of affairs in many applications. It also differs from the foregoing methodologically, since a statistical dimension appears.

What does the item-drawing process tell us about $\underline{\boldsymbol{p}}$? We could simply count the number of times each item comes up in the sampling process, and use it to estimate the probabilities. Good estimates, especially for the smaller probabilities, require inordinately long sampling times, typically much longer than $E[T(\underline{\boldsymbol{p}})]$ (see [29]). We could settle for less, and just inquire about the general shape of the vector: is it close to $\underline{\boldsymbol{e}}/m$? Or to the Zipf distribution? Or to any other familiar distribution? The process $N_1(n)$ that was considered in problem (8) appears interesting in this context. It requires very little overhead in terms of book-keeping, compared with maintaining counters for all items, and is informative. To show this we computed its expected value for several types of $\underline{\boldsymbol{p}}$, for a few values of m and plotted the results, in Fig. 3.

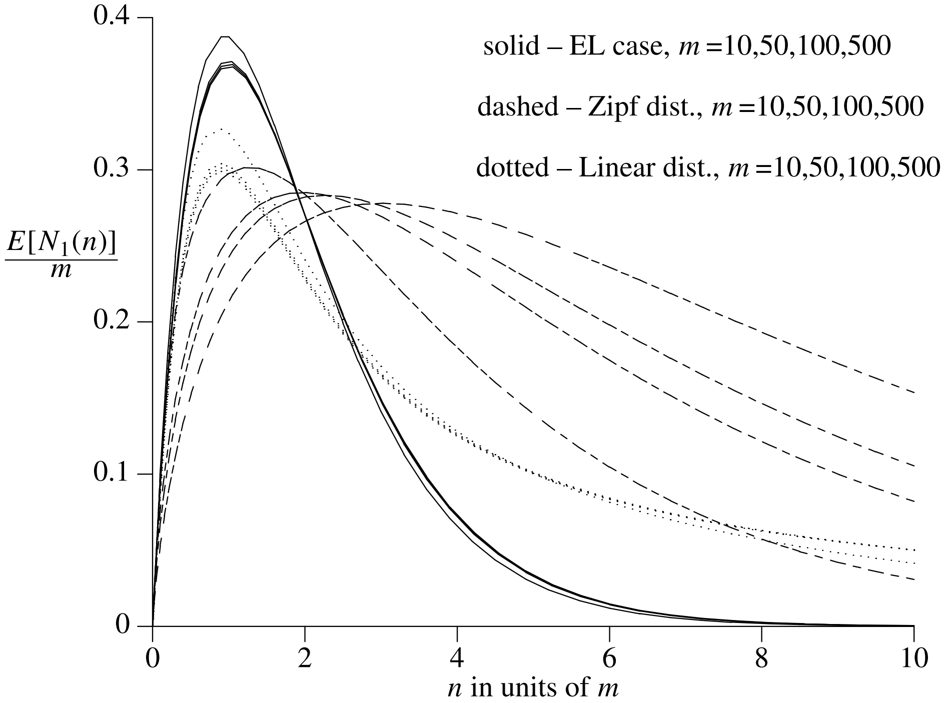

Fig. 3: The fraction of items observed exactly once vs. sample size

For all the distributions, the curve for $m = 10$ peaks higher and sooner than the others. The high dispersion of the curves for the Zipf distribution is the result of the increase of the relative part of small p_i. At least between these families one can distinguish with some experimentation.

Another situation of interest arises when some of the components of $\underline{\boldsymbol{p}}$ might be zero! An example of such situation is the redundancy problem introduced as application (1) in Section 1. Out of the m constraints, let q be hard. The other $r = m - q$ will never show up, regardless of how long the optimizer samples. Since the value of r is not known *a priori*, it is reasonable to ask about a stopping rule which does not depend on the number of sampled items, and in particular – not on its closeness to m. In [40] Robbins makes a remarkable suggestion, attributed there to A.M. Turing: Let $\chi_i(n)$ be a random variable corresponding to item #i, that assumes the value zero if that item has been sampled by the n-th trial, and 1 otherwise. The quantity of interest for the optimizer, that measures "what remains to be done," is $U(\underline{\boldsymbol{p}}, n) = \Sigma p_i \chi_i(n)$ – a random variable that is unobservable, by definition. (Its reciprocal is sometimes called the *resistance* of the process.) Its expected value is easy: $\Pr[\chi_i(n) = 1]$ equals $(1 - p_i)^n$, hence

$$E[U(\underline{\boldsymbol{p}}, n)] = \sum_{i=1}^{m} p_i(1 - p_i)^n. \tag{58}$$

Comparing this with equation (47) we find

$$E[U(\underline{\boldsymbol{p}}, n)] = \frac{1}{n+1} E[N_1(\underline{\boldsymbol{p}}, n+1)]. \tag{59}$$

Since $N_1(\underline{\boldsymbol{p}}, n+1)$ is certainly observable, the optimizer has, at 'time' $n+1$, an unbiased estimate of $U(\underline{\boldsymbol{p}}, n)$. The curves in Fig. 3 should be viewed again in the light of this characterization. To find the variance of $U(\underline{\boldsymbol{p}}, n)$ we compute $E[U^2(\underline{\boldsymbol{p}}, n)]$, as

$$E[(\sum_i p_i \chi_i(n))(\sum_j p_j \chi_j(n))] = \sum_{i=1}^{m} p_i^2(1 - p_i)^n + \sum_{i \neq j} p_i p_j (1 - p_i - p_j)^n. \tag{60}$$

It is easy to find observables that estimate these sums. Let $N_2(\underline{\boldsymbol{p}}, n)$ be the number of items drawn exactly twice in the first n trials, and $N_{(2)}(\underline{\boldsymbol{p}}, n)$ be $\binom{N_1}{2}$, the number of distinct pairs observed once (in any order) in that sequence. Then immediately

$$E[N_2(\underline{\boldsymbol{p}}, n)] = \binom{n}{2} \sum_{i=1}^{m} p_i^2(1-p_i)^{n-2}$$
$$E[N_{(2)}(\underline{\boldsymbol{p}}, n)] = n(n-1) \sum_{i \neq j} p_i p_j (1-p_i-p_j)^{n-2}. \qquad (61)$$

Hence the unbiased estimate

$$\hat{V}[U(\underline{\boldsymbol{p}}, n)] = \frac{N_2(\underline{\boldsymbol{p}}, n+2) + \tfrac{1}{2} N_{(2)}(\underline{\boldsymbol{p}}, n+2)}{\binom{n+2}{2}^2} - \frac{N_1^2(\underline{\boldsymbol{p}}, n+1)}{(n+1)^2}\ .$$

The critical value for the viability of this approach is the variance of the difference $U(\underline{\boldsymbol{p}}, n) - N_1(\underline{\boldsymbol{p}}, n+1)/(n+1)$. It is available from results above, except the expected value of their product. To compute it write $N_1(n) = \Sigma_i \zeta_i(n)$, where the ζ_i are 0–1 variables with the obvious interpretation, and find

$$E\big[U(\underline{\boldsymbol{p}}, n)\, N_1(\underline{\boldsymbol{p}}, n+1)\big] = \sum_{i=1}^{m} p_i^2(1-p_i)^n + \sum_{i \neq j} p_i p_j (1-p_i-p_j)^{n-1}\big((n+1)(1-p_j) - p_i\big). \qquad (62)$$

Estimates for the result are apparent from equation (61) and the similar statistic of the number of pairs with a (2,1) record.

The reader must have realized that this is not a standard estimation procedure; we are "estimating" a random variable rather than a parameter. Hence, in addition to bias, the question of the correlation between this variable and its estimator is of importance. It is not easy to see from equation (62) that the correlation here is frequently negative, but it is – and until n is of the order of the time to completion – quite large. In several experiments its size did not seem to depend much on the shape of $\underline{\boldsymbol{p}}$ except indirectly: it is a slowly decreasing function of the ratio $n/E[T(\underline{\boldsymbol{p}})]$. As this ratio goes to 1, the correlation typically goes from the range (–0.6,–0.4) to about (–0.1,–0.05), and the same rate of decrease persists. At larger values of n it becomes positive (and stays very small). At the same time the variance ratio of $U(\underline{\boldsymbol{p}}, n)$ (and of its estimator) *increases*, hence the estimate should be considered to provide at best order-of-magnitude information only – in spite of its being unbiased!

In a different scenario m is not known, and we inquire about estimating directly the *number* of coupons that have been missed, or the number of new items the sampler may expect to find in the next n' trials, following a sample of size n. We discuss this "prediction" problem below, after introducing our second computational tool.

5. THE RELATED POISSON PROCESS AND PREDICTION

There is a different way of looking at the sampling process, that is suggested by equation (30). Consider m independent Poisson counting processes, with parameters $\{p_i\}$. Since the probabilities sum to one, the expected total number of counts per time unit is one as well. This allows us to establish an equivalence between the 'time' of those processes and the 'time' of the coupon sampling process − which is simply the number of sampled coupons. For example, the probability that each of the Poisson processes produces at least one count by time t is $\Pi_{i=1}^{m}\left(1 - e^{-p_i t}\right)$. Hence the expected time until they all count (at least once) is given by equation (30). Similarly, equation (32) is the expected time to get $m-1$ distinct counts.

Bickel and Yahav show in [5] that these processes are a limiting form of the discrete ones, and use them as asymptotic approximations. Holst shows in [31] a deeper relationship between the two schemes, and obtains a result which is related to Problem (1) above. We show it, using his notation: Let $\{Z_n\}$ denote the time intervals between successive counts, and X_n − the type of the nth arrival. He considers the stopping time $T_{k;m}$, which is when exactly k of the processes have reached—or exceeded—their quota (as in problem (1), process #i has a quota of r_i). Let T_i denote the time until process i fills its quota; it has the Erlang distribution with parameters (r_i, p_i). Now he observes that $T_{k;m}$ is the kth order statistic of $\{T_i\}_{i=1}^m$. If $T_{k;m}$ falls at the $W_{k;m}$th arrival, we have that

$$T_{k;m} = \sum_{j=1}^{W_{k;m}} Z_j, \tag{63}$$

where $W_{k;m}$ and the Z_j are all independent. Now Holst computes the EGF of both sides,

$$E[\exp(x\,T_{k;m})] = E_W\Big[E_Z[\exp\Big(x \sum_{j=1}^{W_{k;m}} Z_j\Big)\Big|W_{k;m}]\Big], \tag{64}$$

and since the interarrival periods are $i.\,i.\,d. \sim \exp(1)$, this gives

$$E[\exp(x\,T_{k;m})] = E[(1-x)^{-W_{k;m}}], \tag{65}$$

whence an immediate relationship follows between the moments of the continuous time $T_{k;m}$ and the (ascending) factorial moments of its discrete version $W_{k;m}$.

The *distribution* of the number of unsampled coupons after n drawings approaches asymptotically, for large m and n, the Poisson distribution with parameter $\Sigma_{i=1}^m \exp(-np_i)$. Holst *et al.* show in [32] an estimate of the rate of convergence to this limit, for $\underline{\boldsymbol{p}}$ which is not 'too far' from the EL case. They prove that when all the probabilities satisfy $c_1/m \le p_i \le c_2/m$, the rate of convergence to this distribution is bound *from above* by $C \cdot \max(m^{-c_1/c_2}, m^{-1/2}\ln m)$, for some constant C.

Poisson Transform

A systematic way to relate the two processes uses the Poisson transform which was introduced in [25] for the EL case, and generalized for unequal probabilities in [28, §7.4]. Let $A(t)$ be a functional over the Poisson processes up to time t, such as a moment of a counter or a probability related to some stopping time, and let A_n be the corresponding functional with respect to the first n samples of the discrete process. The value of $A(t)$ may be computed by conditioning on the number of 'arrivals' during t, since given that n arrivals occurred, they are distributed among the coupon types according to the same underlying multinomial distribution. Hence

$$A(t) = \sum_{n\ge 0} \Pr[n \text{ arrivals during } t]A_n = \sum_{n\ge 0} e^{-t}\,\frac{t^n}{n!}\,A_n. \tag{66}$$

In other words, $A(t)e^t$ is the exponential generating function of $\{A_n\}$. If we can compute the first – the second is immediately available:

$$A_n = n![t^n]A(t)e^t. \tag{67}$$

We have thus another stochastic process, in continuous time, of *independent* processes (unlike the coupon sampling processes, where the discreteness of the time measure introduces dependencies), which provides a handle on the original, less tractable one. For example, since we have $\Pi_i\big(1 - e^{-p_i t}\big)$ for the probability to complete by time t, equation (67) provides $n![t^n]\Pi_i\big(e^{p_i t} - 1\big)$ for the probability of completing within n purchases; viz. equation (41). Many other derivations become startlingly easy this way.

We return now to the question of prediction, introduced at the end of Section 4. First in the Poisson context: The process is observed for a duration T, and the number of times each coupon appears is recorded. We want to compute $\psi(T,t)$, the expected number of coupons that were not observed during T, but will be observed during an additional observation period of t. The contribution of coupon #i to this expectation is

$$\Pr[\text{Coupon } \#i \text{ is not observed in } T \text{ and is observed in } t] = e^{-p_i T}\big(1 - e^{-p_i t}\big). \tag{68}$$

This gives immediately, as in [9],

$$\psi(T,t)=\sum_{i=1}^{m} e^{-p_iT}\Big(1-e^{-p_it}\Big). \tag{69}$$

As $t \to \infty$ we find an estimator for the total number of unobserved types, $\psi(T) = \Sigma_{i=1}^{m} \exp(-p_iT)$. Equation (66) leads to $\exp(T+t)\psi(T,t)$ being the (double) exponential generating function of $\psi(N,n)$. Hence we find the obvious-looking result

$$\psi(N,n)=N!n![T^Nt^n]\sum_{i=1}^{m} e^{T(1-p_i)}\Big(e^t-e^{t(1-p_i)}\Big)=\sum_{i=1}^{m}(1-p_i)^N(1-(1-p_i)^n). \tag{70}$$

Naturally, the collector knows neither m nor the p_i, but the first N observations provide estimates. Specifically, if coupon #i was observed k times, the maximum-likelihood estimate of p_i is k/N. The number of coupons with this estimated reference probability is just the N_k we introduced above. Denoting by n_k the observed value of N_k, equation (70) provides an estimate for the "prediction function" $\psi(N,n)$,

$$\hat{\psi}(N,n)=\sum_{k\ge1} n_k(1-\frac{k}{N})^N(1-(1-\frac{k}{N})^n). \tag{71}$$

Like $\psi(T)$, an estimate for the total number of coupons unseen during the initial sampling is obtained as $n \to \infty$: $\hat{\psi}(N) = \Sigma_{k\ge1} n_k(1-k/N)^N \approx \Sigma_{k\ge1} n_k e^{-k}$.

This procedure appears far odder than estimating $U(\underline{\boldsymbol{p}},N)$, the amount of 'missing probability': even if $U(\underline{\boldsymbol{p}},N)$ were known, the sampling process up to N provides no information whatsoever about the *number* of coupon types among which this unobserved probability is divided. There is no denying the intriguing structure of this problem, though.

The basic idea is apparently due to Good and Toulmin [26]. They used the expression for $E[N_k(T)]$ in the Poisson model,

$$E[N_k(T)]=\sum_{i=1}^{m}\Pr[\text{type } i \text{ arrives } k \text{ times during } T]=\sum_{i=1}^{m} e^{-p_iT}\,\frac{(p_iT)^k}{k!}\,. \tag{72}$$

When this is inserted into a power series development of the second factor of each term in equation (69)—absolute convergence of the sums allows all the manipulations—we get

$$\begin{aligned}\psi(T,t)&=\sum_{i=1}^{m} e^{-p_iT}\sum_{k\ge1}(-1)^{k+1}\frac{(p_it)^k}{k!}\\ &=\sum_{k\ge1}(-1)^{k+1}\Big(\frac{t}{T}\Big)^k\sum_{i=1}^{m} e^{-p_iT}\frac{(p_iT)^k}{k!}=\sum_{k\ge1}(-1)^{k+1}E[N_k(T)]\Big(\frac{t}{T}\Big)^k.\end{aligned} \tag{73}$$

Hence they introduced the following estimator,

$$\hat{\psi}_1(T,t)=\sum_{k\ge1}(-1)^{k+1}n_k\Big(\frac{t}{T}\Big)^k, \tag{74}$$

which has a similar flavor to the estimator shown for $U(\underline{\boldsymbol{p}},n)$. A difficulty this estimator shares with $\hat{\psi}(N,n)$ is the need for $E[N_k(T)]$, since the observed values n_k are very inefficient estimators of these expectations. In practice the situation with $\hat{\psi}_1$ is much worse: it is a polynomial in t, which must lead to serious errors when large values of t are used. It may not even be positive for all t! This estimator, with some corrections, is used in [4] to estimate the number of executions under a regime that imposed imperfect news blackout, and in [17] to estimate the number of words of English that Shakespeare knew, on the basis of his available output. A total of $N = 884647$ words of his oeuvre were counted to obtain the values of n_k. The number of different words among the near million is 31534 (actually, this is the number of 'word types,' a term that counts separately certain variations). Since the estimator (74) does not have a limit, the authors employ a number of ingenious ideas to tame it, and obtain a limit they denote by $\hat{\Delta}(\infty)$. Their conclusion is that "[The] estimate $\hat{\Delta}(\infty) = 35000$ is a reasonably conservative lower bound for the amount of vocabulary Shakespeare knew but did not use." When the same data is plugged in the estimator $\hat{\psi}(884647)$, we obtain the value 6027. If he doubled his output (i.e. $n = N$), $\hat{\psi}(N, N)$ predicts that 3985 new words would have been used. While these numbers appear to us somewhat likelier[2] than the other estimate, a fact due to $\hat{\psi}(N)$ not suffering polynomial growth, it is clearly biased, and possibly heavily so. For example, we need a value for m in $\psi(N)$; our best estimate for it is naturally $\hat{m} = \Sigma n_k$ (this is the maximum-likelihood estimate), missing precisely the number we wish to estimate. One of us proposed in [9] a correction algorithm, to compensate for this bias. We do not know yet what can be claimed about its properties.

It would be obvious from the above that the prediction problem involves serious model-theoretic issues. A good recent survey [12] goes into many of these in greater detail. It also contains an extensive bibliography on the subject, and echoes what is apparently the consensus of the statistical community, that the problem cannot be satisfactorily resolved without additional assumptions. An example is the "Fisher assumption" (following [20], where $\underline{\boldsymbol{p}}$ is assumed to approximate a gamma distribution over the coupon types).

6. CONCLUSION

We have shown that for computations involving CCP-like processes one may reason with various tools. We described first simple operations over languages. This mechanism frequently produces results in terms of integrals that are easy to evaluate, often requiring computation times which are linear in the number of coupon types, whereas exact calculations need exponential time. We then showed how the Poisson transform may be used to obtain certain results for the CCP from a related, much simpler process.

The CCP and several similar problems are considered to be particular cases of the "graph covering problem" (or Markov chain simulation), where an agent is moving from node to node (state to state) and the same questions that arise in the CCP can be asked [45]. However, the CCP is *very* special—corresponding to complete graphs—and the above tools are not suitable for the more general problem – except in highly degenerate cases that can be at least approximated by the CCP. See [2] for a recent reference to these issues.

ACKNOWLEDGMENT

We are grateful to David Aldous for drawing our attention to reference [31], Marcel Neuts for reference [37], anonymous referees for [11,23,36], and to Sheldon Ross for the second proof in problem (9) of Section 3.

REFERENCES

1. Aldous D.: private communication, 1989.
2. Aldous D., Fill J.A.: *Reversible Markov Chains and Random Walks on Graphs,* Book in preparation, expected: 1996.
3. Bardell P.H., McAnney W.H., Savir J.: *Built-in test for VLSI – pseudorandom techniques,* J. Wiley, 1987.
4. Bickel P.J., Yahav J.A.: On Estimating the Number of Unseen Species – How Many Executions Were There? Technical Report No. 43, Dept. of Statistics, UC Berkeley California, June 1985.
5. Bickel P.J., Yahav J.A.: On Estimating the Number of Unseen Species and System Reliability. In S.S. Gupta and J.O. Berger (Eds.): *Statistical Decision Theory and Related Topics IV,* Vol. 2, Springer-Verlag New York, 1988.
6. Boneh A.: "PREDUCE" – A Probabilistic Algorithm Identifying Redundancy by a Random Feasible Point Generator. Chapter 10 in M. Karwan, V. Lotfi, J. Telgen, S. Zionts (Eds.): *Redundancy in Mathematical Programming,* 1983.

[2] This article contains nearly 1250 different words. The 600+ page book [28] contains close to 3800, (probably ≈ 5000 word types).

7. Boneh A.: One Hit-Point Analysis. Unpublished Technical Report, November 1986.
8. Boneh A.: Prediction of the Fault-Detection Curve in Combinatorial Circuits. *IBM Israel Technical Report 88.253,* September 1988.
9. Boneh S., Boneh A., Caron R.J.: Estimating the Prediction Function and the Number of Unseen Species in Sampling with Replacement. Technical Report WMSR-93-07 Department of Mathematics, Windsor University, Ontario Canada. To appear *JASA*, 1996.
10. Boneh S., Papanicolaou V.G.: General Asymptotic Estimates for the Coupon Collector Problem. To appear *J. Computat. Appl. Math.*, 1996.
11. Brayton R.K.: On the Asymptotic Behavior of the Number of Trials Necessary to Complete a Set with Random Selection. *J. Math. Anal. Appl.*, **7**, 31–61 (1963)
12. Bunge J., Fitzpatrick M.: Estimating the Number of Species: a Review. *JASA*, **88**, 364–373 (1993).
13. Caron R.J., Hlynka M., McDonald J.F.: On the Best-Case Performance of Probabilistic Methods for Detecting Necessary Constraints. Technical Report WMSR-88-02, Dept. of Mathematics and Statistics, University of Windsor, Ontario, Canada, February 1988.
14. Caron R.J., McDonald J.F.: A New Approach to the Analysis of Random Methods for Detecting Necessary Linear Inequality Constraints. *Math. Prog.*, **43**, 97–102 (1989).
15. Comtet L.: *Advanced Combinatorics*, D. Reidel, Dordrecht, 1974.
16. David F.N., Barton D.E.: *Combinatorial Chance,* Charles Griffin & Co. London, 1962.
17. Efron B., Thisted R.: Estimating the Number of Unseen Species: How Many Words Did Shakespeare Know? *Biometrika*, **63**, 435–447 (1976).
18. Eldred R.D.: Test Routines Based on Symbolic Logical Statements. *J. of the ACM*, **6**, #3, 33-66 (1959).
19. Feller W.: *An Introduction to Probability Theory and its Applications,* Vol. 1, 3rd Ed. J. Wiley, 1968.
20. Fisher A., Corbet A.S., Williams C.B.: The Relation between the Number of Species and the Number of of Individuals in a Random Sample of an Animal Population. *J. Anim. Ecol.*, **12**, 42–48 (1943).
21. Flajolet P., Gardy D., Thimonier L.: Birthday Paradox, Coupon Collectors, Caching Algorithms and Self-Organizing Search. *Disc. Appl. Math.*, **39**, 207–229 (1992).
22. Gail M.H., Weiss G.H., Mantel N., O'Brien S.J.: A Solution to the Generalized Birthday Problem with Application to Allozyme Screening for Cell Culture Contamination. *J. Appl. Probab.* **16**, 242–251 (1979).
23. Flatto L.: Limit Theorems for Some Random Variables Associated with Urn Models. *Ann. of Probab.*, **10**, 927–934 (1982).
24. Forsythe G.E., Malcolm M.A., Moller C.B.: *Computer Methods for Mathematical Computations.* Prentice-Hall, 1977.
25. Gonnet G.H., Munro J.I.: The Analysis of Linear Probing Sort by the Use of a New Mathematical Transform. *J. of Algo.* **5**, 451–470 (1984).
26. Good I.J., Toulmin G.H.: The number of New Species, and the Increase in Population when a Sample is Increased. *Biometrika*, **43**, 45–63 (1956).
27. Henrici P.: *Applied and Computational Complex Analysis,* Vol. 2, J. Wiley & Sons, 1977.
28. Hofri M.: *Analysis of Algorithms*, Oxford University Press, New York, 1995.
29. Hofri M., Shachnai H.: Self-Reorganizing Lists and Independent References — a Statistical Synergy. *J. of Algo.*, **12**, 533–555 (1991).
30. Holst L.: A Unified Approach to Limit Theorems for Urn Models. *J. Appl. Probab.* **16**, 154–162 (1979).
31. Holst L.: On Birthday, Collectors', Occupancy and Other Classical Urn Problems. *Intern. Stat. Rev.*, **54**, 15–27 (1986).
32. Holst L., Kennedy J.E., Quine M.P.: Rates of Convergence for Some Coverage and Urn Problems Using Coupling. *J. Appl. Probab.* **25**, 717–724 (1988).
33. Janson S.: Limit Theorems for Some Sequential Occupancy Problems. *J. Appl. Probab.* **20**,545–553 (1983).
34. Joag-Dev K., Proschan F.: Birthday Problem with Unlike Probabilities. *American Math. Month.,* **99** 10–12 (1992). Viz. M.V. Hildebrand: The Birthday Problem. *American Math. Month.,* **100**, p.643 (1993).

35. Johnson N.L., Kotz S.: *Urn Models and their Applications. An Approach to Modern Discrete Probability Theory.* John Wiley, New York, 1977.
36. Meilijson I., Newborn M.R., Tenenbein A., Yechiali U.: Number of Matches and Matched People in the Birthday Problem. *Commun. Statist.–Simula. Computa.,* **11**, #3, 361–370 (1982).
37. Neuts M.F., Carson C.C.: Some Computational Problems Related to Multinomial Trials. *Canad. J. of Stats.* Section C, **3**, #2, 235–248 (1975).
38. Newman D.J., Shepp L.: The Double Dixie Cup Problem. *Amer. Math. Month.*, **67**, 58–61 (1960).
39. Rényi A.: Three New Proofs and a Generalization of a Theorem of Irving Weiss. *Magy. Tudo. Akad.*, **7**, 203–213 (1962).
40. Robbins H.E.: Estimating the Total Probability of the Unobserved Outcome of an Experiment. *Ann. Math. Stat.*, **39**, 256–257 (1968).
41. Ross S.M.: Private communication, 1990.
42. Ross S.M.: *Stochastic Processes,* J. Wiley, 1983.
43. Smith R.L., Telgen J.: Random Methods for Identifying Nonredundant Constraints. *Technical Report 81-4, Department of IOE, University of Michigan,* Ann Arbor 1981.
44. Sobel M., Uppuluri V.R.R., Frankowski K.: *Dirichlet Integrals of Type 2 and Their Applications,* Vol. IX in the series *Selected Tables in Mathematical Statistics,* Amer. Math. Soc. 1985.
45. Takács L.: Random Flights on regular graphs. *Adv. Appl. Probab,***16**, 618–637 (1984).
46. Torn A., Zilinskas A.: Global Optimization. In *Lecture Notes in Computer Science* No. 350, Springer-Verlag New York, 1989.
47. Winkler P., Zuckerman D.: Multiple Cover Time. *Random Struct. Alg.,* **9**, 403–411 (1996).