1989

The Coupon-Collector Problem Revisited

Arnon Boneh

Micha Hofri

Report Number:
90-952

The Coupon-Collector Problem Revisited

Arnon Boneh and Micha Hofri
Computer Sciences Department
Purdue University
West Lafayette, IN 47907

CSD-TR-952
February, 1990

THE COUPON-COLLECTOR PROBLEM REVISITED

Arnon Boneh – IOE Department, University of Michigan, Ann Arbor MI 48109-2177
Micha Hofri† – Department of Computer Science, The Technion–IIT, Haifa

May 1989
(Revised January 1990)

ABSTRACT

A standard combinatorial problem calls to estimate the expected number of purchases of coupons needed to complete the collection of all possible m different types. Generalizing this problem, by letting the coupons be obtained with an arbitrary probability distribution, and considering other related processes, the problem has been found to model many practical situations. The usefulness of this model has been seriously hampered by the computational difficulties in obtaining any numerical results concerning moments or distributions. We show, following Flajolet *et al.* [15], that the calculus of generating functions over regular languages may be applied to the problem, answer numerous questions about the sampling process and demonstrate their numerical efficiency. We also present a proof of a long-standing folk-theorem, concerning the extremality of uniform reference probabilities. The paper concludes with a discussion of estimation problems related to the engineering applications of this problem.

1. INTRODUCTION

The Coupon-Collector Problem (CCP) is defined as follows: A set contains m distinct objects: balls in an urn, letters in an alphabet, comics figures taken from the vaults of Disney Productions and sold with chocolate bars... The collector samples from the set *with* replacement. On each trial he has a fixed probability p_i of drawing the i-th object, independently of all past events. Several variables—or processes—are associated with the sequence of trials, all depending on the sample probability vector $\boldsymbol{p} = (p_1, p_2, \cdots, p_m)$:
$X_n(\boldsymbol{p})$ — The item number (or 'name') drawn on the n-th trial.

†Currently at the Department of Computer Science, Purdue University, W. Lafayette, In. 47907.

$T(\boldsymbol{p})$ — The number of the trial that completes the collection.
$T^{(k)}(\boldsymbol{p})$ — The number of the trial that completes the kth full collection.
$T_j(\boldsymbol{p})$ — The number of the trial that completes a sub-collection of size j. Clearly $T(\boldsymbol{p}) = T_m(\boldsymbol{p})$.
$T_C(\boldsymbol{p})$ — The number of the trial that completes a sub-collection C of *specified* objects.
$Y_n(\boldsymbol{p})$ — The number of different items observed in the first n trials.
$N_k(\boldsymbol{p}, n)$ — The number of different items observed *exactly* k times during the first n draws.

Note: The verb 'complete' has always, in the present context, the meaning of 'complete for the first time'. The traditional symbols E and V will be used throughout to denote the expectation and variance of random variables.

The CCP has an extensive past; so extensive, in fact, that even a concise run-down of existing results is well beyond the scope of this report. However, the vast majority of the results we have seen concern the classical problem, where all the coupons have the same probability, m^{-1}, of being drawn. This case is not our main concern. We shall have several occasions below to cite works that are relevant to ours.

The CCP belongs to the family of *Urn problems* [25]. It is a natural framework in which to cast combinatorial questions (and cited as such in [14]). Consequently it has seen numerous applications, though these rarely preserve the ideal simplicity of the original CCP. Here are three examples:

Applications of the CCP

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

A class of algorithms for detecting such constraints, when the feasibility region is convex and of full dimension d, is based on a CCP-like sampling process. One such algorithm is PREDUCE (for Probabilistic REDUCE) suggested by Boneh and Golan [4] and incorporated in an optimization package. Each iteration of PREDUCE consists of generating a ray in random direction, passing through a randomly selected interior feasible point and which hits the boundary of the feasible region. The hit point thus created is a boundary point of the feasible region, and it can be shown that the facet(s) on which it is located belongs with probability one to the hard-constraints set. The algorithm proceeds to generate rays until some stopping rule is satisfied. All the constraints not hit by that time are assumed to be redundant – possibly erroneously. Each such a trial corresponds to drawing one coupon. The number of items is known; however each probability p_i is not known, and is not constant, because the selected interior point is not necessarily the same in all trials: it can be said to be proportional to the *expected* $d-1$ dimensional angle subtended by the corresponding facet, over the set of ray origins. For more details on this set of algorithms see [29]. A direct determination of the non-redundant set is equivalent to computing the probabilities $\boldsymbol{p}$, and is usually very hard, in terms of common representations of constraint sets.

(2) Determining the convex closure of a set of points in R^n.

This problem appears to be related to the first one, but behaves rather differently. A CCP-like determination of the subset of points that span the closure proceeds by generating random $n-1$-dimensional hyperplanes and computing the distances of all points from each. The extreme values (one, if they are all of one sign, or two otherwise) belong to points that are in the desired susbset. Note that if some of the closure hyperplanes contain more than $n-1$ points, some of these points will never be discovered (it is simple to visualize this for $n=2$ with triads (or larger sets) of colinear points: the intermediate ones have a zero probability of being extremally distant from any random line).

(3) The fault-detection (FD) problem in combinatorial circuits.

A combinatorial circuit may 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 (the input vector) produces an output vector on the output set – ostensibly according to the design specifications of the circuit. However, circuits fail sometimes. The standard fault model, "stuck at" [13], 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, and the probability of having more than one fault at any one time—assuming no previous faults—is negligible.

(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 differs from the specified output vector by at least one bit. Correct output may be produced (for certain input vectors) by a faulty circuit. The FD problem consists of finding a list of input vectors—as short as possible, since in critical applications the test is performed very frequently—which detect all the entries in the fault-list.

One approach to the FD problem is to select random input vectors, simulate a faulty circuit and determine the faults detected by that input. 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). Again, searching for such a set of vectors can be viewed as CCP-related: input vectors 'draw' faults at random. There is, however, the added complication that 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 can be estimated by the designer. Further details may be found in [2].

In view of these applications (and others) it is hardly surprising that the literature dealing with the CCP and its ramifications is rich enough. It might be surprising that we found reason to add to it, but we expect the following sections will support the need. Section 2 will survey the

main results we find in the open literature and comment how numerical difficulties have stymied much of the applications of the CCP.

Section 3 brings a relatively unfamiliar device—shuffling of regular languages—and shows its effectiveness in producing numerical values for moments of various variables associated with the CCP. It also leads us to a proof of a property of the CCP that has attained the status of a folk-theorem, but has apparently never been proved: among all possible $\boldsymbol{p}$, the uniform vector produces the shortest expected collection completion times.

Section 4 brings some numerical examples and discusses the computational techniques we used. We then consider various statistical problems when CCP-like processes are employed in practice. As the above sample applications show, the individual probabilities are usually unknown. Even the number of items with non-zero probabilities is not always known to the user. Much work has been done on several aspects of these difficulties. We summarize some of this work and relate it to our methods.

2. RELATED WORK

Texts on probability commonly use the CCP as an example for elementary derivations of expectations. So does Feller (in [14]), who also considers the waiting times between successive increases of the ''observed set''. David and Barton consider in [11] 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 explicit form for the results). Most of the treatments we know of concern the expected time to complete the collection, $E[T(\boldsymbol{p})]$, and its (mainly asymptotic) properties, in the classical case, of stochastically indistinguishable elements. The others will be mentioned during our discussions.

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

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

which is easy to prove from the inclusion-exclusion principle (see [8]) and the relation $E[T(\boldsymbol{p})] = \Sigma_{t\ge 0}\mathrm{Prob}(T(\boldsymbol{p}) > t)$. The earliest source we noticed for equation (1) is [11], who derive it and provide some distributional results.

For the special case of a uniform sampling vector $\boldsymbol{p} = \boldsymbol{e}/m$, where $\boldsymbol{e} = (1, 1, \cdots, 1)$, commonly called the *equally likely* (EL) case, it is possible to obtain more compact expressions, and for higher moments as well, since $T(\boldsymbol{e}/m)$ is representable as a sum of independent geometrically distributed random variables, the parameters of which only depend on their position in the sum – rather than on the particular items that had been sampled. Specifically,

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

The notation H_m stands for the m-th Harmonic number, $\Sigma_{i=1}^{m} 1/i$, and $H_{m-1}^{(2)}$ for $\Sigma_{i=1}^{m-1} 1/i^2$, which converges rapidly to $\zeta(2) = \pi^2/6$.

A. Boneh [5] has obtained a different expression for $E[T(\boldsymbol{p})]$ by considering the different orders in which the coupons may be obtained, and conditioning on that sequence. He finds

$$E[T(\boldsymbol{p})] = \sum_{\boldsymbol{i} \in S(N_m)} P(\boldsymbol{i}) g(\boldsymbol{i}), \tag{3}$$

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. Denoting a particular permutation by $\boldsymbol{i} = (i_1, \cdots i_m)$ he writes

$$P(\boldsymbol{i}) = \frac{\prod_{r=1}^{m} p_r}{\prod_{j=1}^{m}(\sum_{k=j}^{m} p_{i_k})}, \qquad g(\boldsymbol{i}) = \sum_{r=0}^{m-1} \frac{1}{1 - \sum_{k=1}^{r} p_{i_k}}, \tag{4}$$

for the probability of obtaining the collection in the order $\boldsymbol{i}$ and the expected time to complete it in that order, respectively. A convenience of this result is that it often produces ready rough bounds on $E[T(\boldsymbol{p})]$, by computing the function $g(\cdot)$ for two extreme cases: first the most likely order (items are sampled in order of decreasing probability), and secondly, the reverse one – when the 'rarest' item is sampled first, and one ends by finding the item with highest p_i—this can be shown to be the least likely order. The mean must be between these two, but the gap may be substantial and we are hard put to produce from it a tighter bound.

Indeed, as David and Barton comment ruefully, the expressions in equations (1) and (3) are not amenable for numerical evaluation even for moderate m, requiring a large number of operations: on the order of 2^m and $m!$ respectively. Another expression of intermediate complexity is also produced in [8]: 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

$$E[T(\boldsymbol{p})] = E_{N_m}, \tag{5}$$

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

$$E_{\{i_1, i_2, \cdots, i_k\}} = \frac{1}{p_{i_1} + \cdots + p_{i_k}} + \sum_{r=1}^{k} \frac{p_{i_r}}{p_{i_1} + \cdots + p_{i_k}} E_{\{i_1, i_2, \cdots, i_k\} - \{i_r\}}. \tag{6}$$

Brayton obtains in [7] a result equivalent to our equation (40) (the expected time to complete k collections), and the corresponding variance. Since his main concern is in obtaining asymptotic properties, rather than direct computation, he uses a slightly different setup: the $\{p_i\}$ are assumed to be expressible as $\{F(i/n) - F((i-1)/n)\}$; the distribution $F(\cdot)$ is then assumed to admit a density concentrated on [0, 1], that vanishes nowhere, has a finite variation and achieves its minimal value at a finite number of isolated points. The asymptotic properties turn out to hinge on this "minimum set".

The reference [8] reports on an effort to show the intuitively appealing conjecture for the CCP, that $E[T(\boldsymbol{p})]$ is minimized over all probability vectors in the EL case. This conjecture has been part of the folklore for quite some time [1], but we have not seen it proven anywhere. A reason for the interest in this fact beyond the mere mathematical one is that it would give a natural (and easily computable, for a change) yardstick by which to judge the relative difficulty of CC problems. It is quite easy to show—at least from some of the expressions that were obtained for $E[T(\boldsymbol{p})]$—that $\boldsymbol{e}/m$ is a stationary point. The authors there also show—through equation (1)—that $\boldsymbol{e}/m$ is a strong local minimum. But the proof that it is a global minimum remained elusive. The difficulty is that these expressions are not convex in the components of $\boldsymbol{p}$. A possible way is then to show that $E[T(\boldsymbol{p})]$ *is* convex on the sheet $\Sigma p_i = 1$; it appears there is no way to do this uniformly, for all m. The authors indeed managed to do this for $m \le 6$.

The references [12] and [28] contain accounts of other fascinating questions, of statistical nature, concerning the coupon collecting process.

The most detailed treatment of CCP-related questions is [30]. The authors' starting point is not any specific problem actually, but rather the answer: they consider a few parametrized families of the so-called 'Dirichlet integrals of type 2'. They then show a very large and rich collection of sampling stopping-time problems, the solutions of which can be expressed in terms of these integrals, and the CCP problem falls squarely in that domain. The reference [30] 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 all items have either the same – or at most two different values for the selection probabilities. In order to treat non-uniform sampling probability vectors, the authors provide Taylor expansions of the basic integrals at the equally-likely 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.

Computational Difficulty

We have commented before on the huge computational effort required to obtain the expected value $E[T(\boldsymbol{p})]$. It appears that except for the EL case, only rarely—and then, for rather small m—are any expectations or probabilities explicitly calculated. Often in such situations one tries to use asymptotic results: the precision is often sufficient in practice, and it is of course at large sizes of the problem where at one and the same time numerical difficulties are at their worst, and asymptotic methods at their best. This works fine when $\boldsymbol{p}$ can be characterized by one or two parameters, but in more general cases, where the components of $\boldsymbol{p}$ do not satisfy any convenient relation (beyond summing to 1, that is) no asymptotic results seem forthcoming. The situation seemed so bad that a researcher in an area that applies the CCP was moved to say that once m exceeds 30 or so distinct values, it is immaterial whether one knows the selection probabilities or not – one can compute nothing with them anyway. We shall show, in the next section, that this is definitely not the case: we can routinely compute expectations and probabilities for thousands of items and more, with the effort (for most of the computations) roughly linear in m and essentially independent of the probability values.

3. A NEW APPROACH

Recently, Flajolet et al. presented in [15] a new approach, or rather, a novel point of view of the CCP, which resulted in a computationally superior expression for the expected duration of the CC activity, which they obtained as well. We show both below. Their approach may be viewed in a wider context, as a way to compute probabilities for various variables defined in terms of sequences of independent sampling from a finite population. The same approach was briefly mentioned earlier by Comtet in [10]. In this framework there is a natural way to answer more detailed questions about the sampling process.

After showing the outlines of the method, its application will be illustrated by posing and solving a sequence of such questions. In this section we only provide analytic expressions. The power and utility of the method lie largely—as the authors in [15] observe—in the amenability of these to numerical evaluation. Computational techniques for this purpose is the topic of Section 4.

We collect here the questions, so that their inter-relationships will be easier to perceive:

(1) In a sample of length n, what is the probability that item *#i* occurs at least r_i times? In the CCP the r_i are specialized to be all 1.

(2) The same question as (1), but now we ask only about a subset of the items, $C \subset A$.

(3) What is the probability of finding in a sample of size n at least r items repeated each at least k times?

(4) What is the expected number of different items drawn in a sample of size n?

(5) What is the expected time until $j \le m$ different coupons have been sampled?

(6) What is the expected time until $j \le m$ different coupons have been sampled at least k times each?

(7) What is the expected time until $j \le m$ *specified* different coupons have been sampled?

(8) How many coupons will be sampled exactly r times, or more than r times, before $T(\boldsymbol{p})$?

(9) Is the expected time a good estimate of the required time? – we want more information about the distribution of $T(\boldsymbol{p})$, so we can answer the questions: with what probability will $T(\boldsymbol{p})$ exceed $E[T(\boldsymbol{p})]$ by a certain fraction, or in still another way: how long do we have to sample to obtain all coupons with a probability exceeding α?

(10) What is the distribution of $N_1(\boldsymbol{p},n)$, defined as the number of coupons observed exactly once in the first n trials? In the next section we explain the significance of this variable.

This is what we need in order to solve the above:

Strings Over A Finite Alphabet

Let $A = \{a_1, \cdots, a_m\}$ be an *alphabet*. The set of all finite words of letters from A is denoted by A^*. A subset of A^* is called a *language*, and a word in a language will be generically denoted by w. The letter a_j is associated with the probability p_j, and a word w is assumed to

carry two types of weights: one is the standard additive weight, chosen here to be the size of the word in letters, and denoted by $|w|$. The second is a "probabilistic weight" that equals the product of the probabilities of its letters: $\pi(w) \equiv \Pi_{a_j \in w} p_j$. Define for a language L the following probability generating functions (pgf):

$$\phi_L(z) \equiv \sum_{w \in L} \pi(w) z^{|w|}, \qquad \hat{\phi}_L(z) \equiv \sum_{w \in L} \pi(w) \frac{z^{|w|}}{|w|!}. \tag{7}$$

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

$$\phi_L(z) = \int_{t \geq 0} \hat{\phi}_L(zt) e^{-t} dt. \tag{8}$$

These are not the usual pgf's used in probability theory for the distributions of (discrete) random variables, but observe that $[z^n]\ \phi_L(z)$ is the probability that a random word of size n, from A^*, is in the language L. This formalism does not support directly the notion of a random word of arbitrary size.

We define two operations on languages: concatenation and shuffling, which are used extensively below; they expose two properties of the pgf's we defined and exhibit the need for both types of functions.

1. The *concatenation* of two functions, L_1 and L_2, is a language L, written as $L=L_1L_2$, such that each word of L is formed by concatenating a word from L_2 to a word from L_1, to form $w=w_1.w_2$. The operation is "well-defined" iff it has the property of unique factorization: for each $w \in L$ there exists a unique pair w_1, w_2, such that $w_i \in L_i$, and $w=w_1.w_2$.

Proposition 1: If the operation $L=L_1L_2$ is well-defined, then

$$\phi_L(z) = \phi_{L_1}(z)\phi_{L_2}(z). \tag{9}$$

2. 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. This recursive definition can be formally expressed as follows:

$$\begin{aligned} & w \circ \varepsilon = \varepsilon \circ w = w, && (\varepsilon \text{ is the null word.}) \\ & a.w_1 \circ b.w_2 = a.(w_1 \circ b.w_2) \cup b.(a.w_1 \circ w_2) && a, b \in A, \quad w_i \in L_i, \\ & L_1 \circ L_2 = \bigcup_{\substack{w_1 \in L_1 \\ w_2 \in L_2}} w_1 \circ w_2. \end{aligned} \tag{10}$$

The operation is "well-defined" for languages that use disjoint subsets of A – that is, employ different alphabets. When this is the case, it is straightforward to show that

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

$$\hat{\phi}_L(z) = \hat{\phi}_{L_1}(z)\hat{\phi}_{L_2}(z). \tag{11}$$

All our applications of this tool have the following format: the statement of the problem is

reinterpreted—sometimes this is straightforward and occasionally involves intermediate steps—as a specification of a language H from A^*. The words of H are shown to be constructible from simple components by concatenation and shuffling. The building-blocks will be such that their pgf's are easy to compute directly, and Propositions 1 and 2 will provide the pgf's of H, from which we shall 'read off' the desired answers. We show here one such construction that is used extensively below:

Define a k-hit as the occurrence of a letter k times (or more) in a word. We shall see later that it is useful to be able to compute the probability that a random word of a specified size has k-hits for exactly q distinct letters.

The basic construct is the following one: Let $H_{q,k}$ be a language with only such words in which exactly q letters recur at least k times, and the other $m-q$ letters appear at most $k-1$ times. Then introduce the following notation for uni-letter languages:

$$a^{<k} = \{\varepsilon, a, a^2, \ \cdots \ , a^{k-1}\}, \qquad a^{\geq k} = a^k \cdot a^*, \tag{12}$$

where a^2 is shorthand for the word aa, etc., and a^* is the supremum of all $a^{<k}$. The crux of the approach is that this notation provides the following expression for all the words of $H_{q,k}$:

$$H_{q,k} = \bigcup_{I,J} \left(a_{i_1}^{\geq k} \circ a_{i_2}^{\geq k} \circ \ \cdots \ \circ a_{i_q}^{\geq k}\right) \circ \left(a_{j_1}^{<k} \circ a_{j_2}^{<k} \circ \ \cdots \ \circ a_{j_{m-q}}^{<k}\right), \tag{13}$$

where the union is over all two-set partitions of A, indexed by I and J: $I = \{i_1, \ \cdots \ , i_q\}$, $J = \{j_1, \ \cdots \ , j_{m-q}\}$, $I \cap J = \varnothing$, $I \cup J = N_m$. This may seem merely a complicated way to repeat the verbal specification of $H_{q,k}$, but since the exponential pgf's of the constituent elements are straightforward to write, we shall obtain immediately that of $H_{q,k}$: Let $e_k(z)$ denote the incomplete exponential function

$$e_k(z) = \sum_{i=0}^{k} \frac{z^i}{i!}, \tag{14}$$

then, since a^* contains exactly one word of each size $j \geq 0$, the desired exponential pgf's 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, summing over products of these exponential pgf's, we obtain for the exponential pgf of $H_{q,k}$:

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

This ungainly sum allows for a more compact representation; we need for it the notation $[x^r]f(x)$, for the coefficient of x^r in the power series development of $f(x)$:

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

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—it is more convenient to use the ordinary pgf, and with the Laplace-

Borel Transform we find

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

When all the probabilities are equal, at $1/m$, the expressions are naturally much neater. The summation in equation (16) is over $\binom{m}{q}$ identical terms, yielding

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

and similarly for the ordinary pgf.

The utility of the the machinery outlined above will be now demonstrated in providing answers to problems (1) through (10).

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

The samples satisfying this requirement make up a "language" in which each word contains the letter a_i at least r_i times. Denote this language by $H(\boldsymbol{p}, \boldsymbol{r})$. Then in analogy with equation (13)

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

where the $a_i^{\ge r_i}$ are defined in equation (12). The exponential egf of the collection of uni-letter words $a_i^{\ge r_i}$ is $e^{p_i z} - e_{r_i-1}(p_i z)$, with $e_{-1}(\cdot) \equiv 0$; hence

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

and the desired probability is given by

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

As is usually the case in these problems, there is no ready explicit form for this coefficient. For moderate m and n, its numerical value may be obtained using the Cauchy integral formula (we remark that $\hat{\Phi}(\boldsymbol{p}, \boldsymbol{r}; z)$ is an entire function).

(2) *The probability of drawing coupon #i at least r_i times in n trials,* $i \in C \subset A$

This problem is very similar to problem (1), with no requirements on the elements of $A - C$, which is equivalent to setting there $r_i = 0$ for $a_i \in A - C$. Alternatively, The language we need can be written (with a suitable notation for the indexes of the elements of C)

$$H(\boldsymbol{p}, \boldsymbol{r}, C) = \left(a_{(1)}^{\ge r_{(1)}} \circ a_{(2)}^{\ge r_{(2)}} \circ \cdots \circ a_{(k)}^{\ge r_{(k)}}\right) \circ (A - C)^*, \tag{23}$$

where $a_{(i)} \in C$, $1 \le i \le k = |C|$. The exponential pgf of $(A - C)^*$ is $e^{(1-P_C)z}$, where $P_C = \sum_{i \in C} p_i$. The desired probability is analogously produced by

$$\mathrm{P}(\boldsymbol{p}, \boldsymbol{r}, C; n) = n![z^n]e^{z(1-P_C)}\prod_{i=1}^{k}\left(e^{p_{(i)}z} - e_{r_{(i)}-1}(p_{(i)}z)\right). \tag{24}$$

Naturally this may be obtained from equation (22) when $r_{(i)} = 0$ is inserted there, for $1 \le i \le k$.

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

A k-hit is defined above as the occurrence of a letter at least k times in a word. We also write there the exponential pgf of the language $H_{r,k}$, which contains words with *exactly* r k-hits – these are equations (16) and (17). The construction is indeed similar to the one above, but 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? The temptation to use the language

$$H(\boldsymbol{p}, r, k) = \bigcup_{I:\ |I|=r}\left[\left(a_{i_1}^{\ge k} \circ a_{i_2}^{\ge k} \circ \cdots \circ a_{i_r}^{\ge k}\right) \circ (A-I)^*\right], \tag{25}$$

should be resisted, since many words are repeated in this specification. The way to follow is to use equation (17). Define $Y_n(\boldsymbol{p}, k)$ as the number of k-hits in a word of size n. Then

$$\begin{aligned}\mathrm{Prob}[Y_n(\boldsymbol{p}, k) = j] &= n![z^n u^j]\prod_{i=1}^{m}\left(e_{k-1}(p_i z) + u\left(e^{p_i z} - e_{k-1}(p_i z)\right)\right) \\ &= [z^n u^j]\int_{t\ge 0}\prod_{i=1}^{m}\left(e_{k-1}(p_i zt) + u\left(e^{p_i zt} - e_{k-1}(p_i zt)\right)\right)e^{-t}dt,\end{aligned} \tag{26}$$

where we have used in the last line the Laplace-Borel transform. The required answer is then $\Sigma_{r\le j\le m}\mathrm{Prob}[Y_n(\boldsymbol{p}, k) = j]$. There does not seem to be any essentially simpler way of expressing this truly complex combinatorial quantity. Even in the EL case, while it looks simpler, the computational effort is essentially the same.

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

This problem calls for the evaluation of $E[Y_n(\boldsymbol{p}, 1)]$. The answer is available from elementary considerations (see e.g. [9]), but it should be instructive to use the present apparatus to recapture it. Since equation (26) gives the probabilities, we have

$$\begin{aligned}E[Y_n(\boldsymbol{p}, 1)] &= \sum_{j=1}^{n} jn![z^n u^j]\prod_{i=1}^{m}\left(1 + u\left(e^{p_i z} - 1\right)\right), \\ &= n![z^n]\frac{\partial}{\partial u}b(u, z)\Big|_{u=1},\end{aligned} \tag{27}$$

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

$$\frac{\partial}{\partial u}b(u, z) = b(u, z)\sum_{i=1}^{m}\frac{e^{p_i z} - 1}{1 + u\left(e^{p_i z} - 1\right)}, \tag{28}$$

and at $u = 1$

$$\frac{\partial}{\partial u}b(1,z) = b(1,z)\sum_{i=1}^{m}\frac{e^{p_i z}-1}{e^{p_i z}} = e^{z}\sum_{i=1}^{m}(1-e^{-p_i z}). \tag{29}$$

Finally

$$E[Y_n(\boldsymbol{p},1)] = n!\sum_{i=1}^{m}(\frac{1}{n!} - \frac{(1-p_i)^n}{n!}) = \sum_{i=1}^{m}[1-(1-p_i)^n], \tag{30}$$

as we should expect. In Section 4 we consider some properties of this value.

(5) The expected time to draw a sub-collection of size j [15]

Consider the following set of equalities:

$$E[T_j(\boldsymbol{p})] = \sum_{n\ge 0}\mathrm{Prob}(T_j(\boldsymbol{p}) > n) = \sum_{n\ge 0}\mathrm{Prob}(Y_n(\boldsymbol{p},1) < j). \tag{31}$$

The second equality results from the fact that the two compound events $\{T_j(\boldsymbol{p}) > n\}$ and $\{Y_n(\boldsymbol{p},1) < j\}$ consist of precisely the same sequences of trials, and hence have the same probability. Now we prepare to use equation (26):

$$\sum_{n\ge 0}\mathrm{Prob}(Y_n(\boldsymbol{p},1) < j) = \sum_{n\ge 0}\sum_{r=0}^{j-1}\mathrm{Prob}(Y_n(\boldsymbol{p},1)=r) = \sum_{r=0}^{j-1}\{\sum_{n\ge 0}\mathrm{Prob}(Y_n(\boldsymbol{p},1)=r)\}. \tag{32}$$

Armed now with the necessary ingredients, equations (26) and (18) give, specialized to $k=1$

$$E[T_j(\boldsymbol{p})] = \sum_{r=0}^{j-1}\sum_{n\ge 0}\mathrm{Prob}[Y_n(\boldsymbol{p},1)=r] = \sum_{r=0}^{j-1}[u^r]\int_{t\ge 0}\prod_{i=1}^{m}(1+u(e^{p_i t}-1))e^{-t}dt. \tag{33}$$

Since the integrand is an m-degree polynomial in u, the case $j=m$ 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:

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

and since $\exp(-t) = \exp(-\Sigma p_i t)$, we further simplify to

$$E[T(\boldsymbol{p})] = \int_{t\ge 0}[1-\prod_{i=1}^{m}(1-e^{-p_i t})]dt. \tag{35}$$

We have thus obtained a computationally convenient form for the expected duration of $T(\boldsymbol{p})$.

Equation (35) has a particularly simple form in the particularly simple EL case of *uniform* probability vector:

$$E[T(\boldsymbol{e}/m)] = \int_{t\ge 0}[1-(1-e^{-t/m})^{m}]dt. \tag{36}$$

Interestingly enough, an integral nearly identical with the one in equation (36) (with the transformation of the integration variable to $x=e^{-t}$) may be found in an unpublished report by the authors of [8]. Observe, moreover, that the equality of the integral in (35) to the right-hand

side of equation (1) is straightforward.

The computational advantage of this integral over the sums we encountered in 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 (we shall discuss this below), but the function is smooth enough for an integration routine with locally adaptive step-size to compute it with several hundreds of function evaluations, under very stringent accuracy requirements.

There is interest also in finding the time to 'almost complete' the sampling, that is, for values of j that are very close to m, and the same approach that led to equation (35) applies, with a somewhat heavier price. Thus, for example, we find

$$E[T_{m-1}(\boldsymbol{p})] = \int_{t\geq 0}[1 - \sum_{j=1}^{m}(p_j + e^{-p_j t}(1-p_j))\prod_{\substack{i=1\\ i\neq j}}^{m}(1 - e^{-p_i t})]dt. \tag{37}$$

Note that with some care the computations of these quantities can be still kept to be essentially linear in m.

Remark: We found it instructive to consider the two quantities, computed in problems (4) and (5), side by side. Both may be viewed as functions in the (sampling duration, number of captures) plane, but with the relation of independent/dependent variables reversed. Consider these coordinates as providing the abscissa and ordinate, respectively, as shown in the generic Fig. 1.

The engineering (and statistical) significance of the two functions are also very different: The "detection curve", given by equation (30), shows the expected number of detected items as a function of the sampling duration; we find it more suggestive to think of the fraction of the

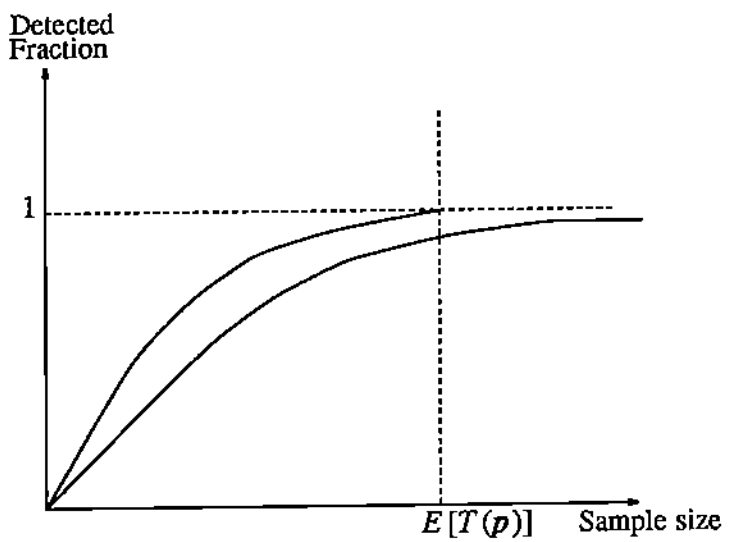

Fig. 1: Detection–temporal relationships

items that have been captured, rather than their number. Then the "sample duration curve", computed via equation (33), gives the expected duration required to complete a specified fraction. Equations (35) and (37) 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 abscissa), while the second one terminates in the point $(E[T(\boldsymbol{p})], 1)$. Another way to distinguish the two curves is to consider how they compare with individual experiments. Thus points of the first curve represent average of samples scattered along the ordinate, while the second one features averages of horizontal scatter. We conjecture the duration curve to lie always (that is, for any $\boldsymbol{p}$) entirely above the detection curve.

(6) The expected time to draw a size j sub-collection k times [7]

For $j=m$, $k=2$, in the EL case, this problem has been known as the "double Dixie cup problem" (after the product that carried the collected coupons). Holst provides an answer for the EL case in [23], and comments on earlier derivations of that result, extending to 1960, all considering complete collections (i.e. $j=m$), and—except [7]—the EL case. Other references of interest (all essentially concerned with limiting properties of $E[T(\boldsymbol{e}/m)]$ and variations thereof) are [16], [26] and [27].

The required expected value is obtained exactly as for Problem (5):

$$E[T_j^{(k)}(\boldsymbol{p})] = \sum_{n\ge 0} \text{Prob}(T_j^{(k)}(\boldsymbol{p}) > n) = \sum_{n\ge 0} \text{Prob}(Y_n(\boldsymbol{p}, k) < j). \tag{38}$$

The same approach that produced equations (33) and (35) will now provide

$$E[T_j^{(k)}(\boldsymbol{p})] = \sum_{r=0}^{j-1} \sum_{n\ge 0} \text{Prob}[Y_n(\boldsymbol{p}, k) = r] = \sum_{r=0}^{j-1} [u^r] \int_{t\ge 0} \prod_{i=1}^{m} (1 + u(e^{p_i t} - e_{k-1}(p_i t)))e^{-t}dt, \tag{39}$$

and

$$E[T^{(k)}(\boldsymbol{p})] = \int_{t\ge 0} [1 - \prod_{i=1}^{m} (1 - e^{-p_i t} e_{k-1}(p_i t))]dt. \tag{40}$$

Computing this expected time is of roughly the same difficulty as that of $E[T(\boldsymbol{p})]$, and increases sub-linearly with k. Equation (40) is precisely the result already obtained in [7]. Brayton also considers asymptotic results (as $m\to\infty$); his results there depend on a particular model he chose to generate $\boldsymbol{p}$, so we shall not go into them in any detail, except to mention that in the EL case he obtains $E[T^{(k)}(\boldsymbol{e}/m)] \approx m[\log m + (k-1)\text{loglog}\, m + C_k]$.

(7) *Expected time to draw a specified sub-collection of size j*

While the formulation appears rather different than problem (5), and may be seen as a significant generalization, it can be handled almost identically. Let the required subset be $C \in A$. As above,

$$E[T_C(\boldsymbol{p})] = \sum_{n\ge 0}\text{Prob}[T_C(\boldsymbol{p}) > n] = \sum_{n\ge 0}\text{Prob}[W_n(C,\boldsymbol{p}) < j] = \sum_{k=0}^{j-1}\{\sum_{n\ge 0}\text{Prob}[W_n(C,\boldsymbol{p}) = k]\}, \quad (41)$$

where $W_n(C,\boldsymbol{p})$ is the number of coupons from C observed in a sequence of n drawings. Still as before, except that different words contribute and therefore we construct a different language: $H_k(C)$ is a language containing only words in which exactly k items of C appear. Using equations (12) and (13) it can be written as

$$H_k(C) = \bigcup_{I\subset C}\left(a_{i_1}^{\ge 1}\circ a_{i_2}^{\ge 1}\circ\ \cdots\ \circ a_{i_k}^{\ge 1}\right)\circ(A-C)^*, \quad (42)$$

where the union is over all $I=\{a_{i_1},\ \cdots\ ,a_{i_k}\}\subset C$, so that $H_k(C)$ has the exponential pgf

$$\hat{\phi}_{k,C}(z) = \sum_{I:\,|I|=k}\prod_{i\in I}(e^{zp_i}-1)e^{z(1-P_C)}, \quad (43)$$

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 (35) provide

$$\hat{\phi}_{k,C}(z) = [u^k]\prod_{i=1}^{j}[1+u(e^{zp_{(i)}}-1)]e^{z(1-P_C)}, \quad (44)$$

and for the expectation we find

$$\begin{aligned} E[T_C(\boldsymbol{p})] &= \sum_{k=0}^{j-1}\phi_{k,C}(1) = \sum_{k=0}^{j-1}\int_{t\ge 0}[u^k]\prod_{i=1}^{j}[1+u(e^{tp_i}-1)]e^{t(1-P_C)}e^{-t}dt, \\ &= \int_{t\ge 0}[1-\prod_{i\in C}(1-e^{-p_{(i)}t})]dt, \end{aligned} \quad (45)$$

in complete analogy with equation (35).

(8) *The distribution of r-hits during* $T(\boldsymbol{p})$

Denote by $M_r(\boldsymbol{p})$ the number of coupons observed at least r times once $T(\boldsymbol{p})$ is over. Words that contribute to $\text{Prob}(M_r(\boldsymbol{p})=k)$ terminate with a letter appearing for the first time and completing the collection. Specifying this letter as a_l, the corresponding prefixes form the language $L_{r,k}(l)$ that has the structure

$$L_{r,k}(l) = \bigcup_{I\subset A-\{l\}}\left(a_{i_1}^{\ge r}\circ a_{i_2}^{\ge r}\circ\ \cdots\ \circ a_{i_k}^{\ge r}\right)\circ\left(a_{j_1}^{+<r}\circ a_{j_2}^{+<r}\circ\ \cdots\ \circ a_{j_{m-k-1}}^{+<r}\right), \quad (46)$$

where $I=\{a_{i_1},\ \cdots\ ,a_{i_k}\}$, $a_{j_c}\in A-I-\{l\}$, for $1\le c\le m-k-1$ and $a^{+<r}$ is a language consisting of words of sizes between 1 and $r-1$. $L_{r,k}(l)$ has then the exponential pgf

$$\begin{aligned} \hat{\lambda}_{r,k}(l;z) &= \sum_{I}\prod_{i\in I}(e^{zp_i}-e_{r-1}(zp_i))\prod_{j\in A-I-\{l\}}(e_{r-1}(zp_j)-1). \\ &= [u^k]\prod_{\substack{i=1\\ i\ne l}}^{m}[e_{r-1}(zp_i)-1+u(e^{zp_i}-e_{r-1}(zp_i))]. \end{aligned} \quad (47)$$

and the ordinary pgf

$$\lambda_{r,k}(l;z) = [u^k] \int_{t\ge 0} \prod_{\substack{i=1 \\ i\ne l}}^{m} \left[e_{r-1}(ztp_i) - 1 + u\left(e^{ztp_i} - e_{r-1}(ztp_i)\right)\right] e^{-t} dt. \tag{48}$$

The complete words that contribute to $\mathrm{Prob}(M_r(\boldsymbol{p}) = k)$ are formed by concatenating the above prefixes with a_l. The letter has the ordinary pgf $p_l z$, and hence, by Proposition 1 and the observation that the probabilistic weight of each word is actually equal to the probability of obtaining the corresponding sample,

$$\mathrm{Prob}(M_r(\boldsymbol{p}) = k) = \sum_{l=1}^{m} p_l [u^k] \int_{t\ge 0} \prod_{\substack{i=1 \\ i\ne l}}^{m} \left[e_{r-1}(tp_i) - 1 + u\left(e^{tp_i} - e_{r-1}(tp_i)\right)\right] e^{-t} dt. \tag{49}$$

For example, the probability that the maximal frequency obtained by any coupon does not exceed $r-1$, is given by the readily computable expression

$$\mathrm{Prob}(M_r(\boldsymbol{p}) = 0) = \sum_{l=1}^{m} p_l \int_{t\ge 0} e^{-t} \prod_{\substack{i=1 \\ i\ne l}}^{m} \left(e_{r-1}(tp_i) - 1\right) dt. \tag{50}$$

Also, the expected number of such multiplicities is given by

$$E[M_r(\boldsymbol{p})] = \sum_{k=1}^{m-1} \sum_{l=1}^{m} k p_l [u^k] \int_{t\ge 0} \prod_{\substack{i=1 \\ i\ne l}}^{m} \left[e_{r-1}(tp_i) - 1 + u\left(e^{tp_i} - e_{r-1}(tp_i)\right)\right] e^{-t} dt. \tag{51}$$

This integral can be simplified to some extent. Since $\Sigma k[u^k]f(u) = f'(1)$, we get

$$E[M_r(\boldsymbol{p})] = \int_{t\ge 0} \sum_{l=1}^{m} p_l e^{-tp_l} \sum_{\substack{j=1 \\ j\ne l}}^{m} \left[1 - e^{-tp_j} e_{r-1}(tp_j)\right] \prod_{i\ne j,\, l} \left(1 - e^{-tp_i}\right) dt. \tag{52}$$

Changing the order of summation and integrating by parts:

$$E[M_r(\boldsymbol{p})] = m - \int_{t\ge 0} \sum_{l=1}^{m} p_l e^{-tp_l} \frac{(p_l t)^{r-1}}{(r-1)!} \prod_{\substack{i=1 \\ i\ne l}}^{m} \left(1 - e^{-tp_i}\right) dt. \tag{53}$$

The language defined in equation (46) can be used to evaluate directly the expectation of $T(\boldsymbol{p})$, rather than the circuitous way that led us to equation (35). It results however in a computationally less-efficient expression for the expectation.

(9) *On the distribution of* $T(\boldsymbol{p})$.

Actually we have obtained already expressions for the distribution of $T(\boldsymbol{p})$: from the discussion leading to $E[T(\boldsymbol{p})]$ – see equation (31), we find for the tail probabilities of $T(\boldsymbol{p})$, specializing equation (26) to $k=1$, that

$$\mathrm{Prob}[T(\boldsymbol{p}) > n] = \mathrm{Prob}[Y_n(\boldsymbol{p}, 1) < m] = \sum_{r=0}^{m-1} n! [z^n u^r] \prod_{i=1}^{m} \left(1 + u\left(e^{p_i z} - 1\right)\right). \tag{54}$$

The same treatment that led to equation (35) also provides

$$\text{Prob}[T(\boldsymbol{p}) > n] = n![z^n]\Big(e^z - \prod_{i=1}^{m}(e^{p_i z} - 1)\Big) = 1 - n![z^n]\prod_{i=1}^{m}(e^{p_i z} - 1). \tag{55}$$

which is as simple as we could ask for. The last equality is also a specialization of equation (22).

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

$$I_2 = \sum_{n\ge 1} n\text{Prob}[T(\boldsymbol{p}) > n] = \sum_{n\ge 1} n\Big(1 - n![z^n]\prod_{i=1}^{m}(e^{p_i z} - 1)\Big). \tag{56}$$

The second one yields

$$\begin{aligned} I_2 = \sum_{n\ge 1} n\text{Prob}[T(\boldsymbol{p}) > n] &= \int_{t\ge 0} \sum_{j=1}^{m} p_j t\Big[1 - \prod_{\substack{i=1\\ i\ne j}}^{m}(1 - e^{-p_i t})\Big]dt. \\ &= \int_{t\ge 0} \sum_{j=1}^{m} p_j t\Big[1 - \frac{\Pi(t)}{1 - e^{-p_j t}}\Big]dt, \qquad \Pi(t) \equiv \prod_{i=1}^{m}(1 - e^{-p_i t}). \end{aligned} \tag{57}$$

Then $V[T(\boldsymbol{p})] = 2I_2 + E[T(\boldsymbol{p})](1 - E[T(\boldsymbol{p})])$. Both expressions for I_2 are cumbersome to compute; the first one is of the same type as equation (35): it could be decomposed to something like equation (1), involving $O(2^m)$ terms, most of which would be vanishingly small and hard to estimate. The second would be – numerically – less troublesome. It is very similar to the integral required for $E[T(\boldsymbol{p})]$, but in order to keep the computation time linear in m, we have to separate the evaluation of the product, as shown there.

(10) *On the distribution of* $N_1(n)$.

Here we are interested in the number of letters that occur exactly once in a word of size n. The words that contribute to $\text{Prob}(N_1(n) = r)$ are similar to those we constructed in problem (3):

$$H(\boldsymbol{p}, r) = \bigcup_{I:\ |I|=r}\Big[\big(a_{i_1} \circ a_{i_2} \circ \cdots \circ a_{i_r}\big) \circ (A - I)^{\neq 1}\Big], \tag{58}$$

$$\hat{\Phi}_r(z) = \sum_{I} \prod_{i\in I} z p_i \prod_{j\notin I}(e^{z p_j} - z p_j), \tag{59}$$

and hence

$$\text{Prob}[N_1(\boldsymbol{p}, n) = r] \;=\; n![z^n u^r]\prod_{i=1}^{m}\big(e^{z p_i} + (u - 1) z p_i\big). \tag{60}$$

In this case as well, numerical evaluations are not simple, and unless the components of $\boldsymbol{p}$ have an analytically tractable form one is unlikely to obtain even useful asymptotic estimates. The expectation is straightforward to obtain either from equation (60) or from elementary considerations, and equals

$$E[N_1(\boldsymbol{p}, n)] = n\sum_{i=1}^{m} p_i(1-p_i)^{n-1}. \tag{61}$$

Sobel et al. in [30] bring up a few more questions that are of interest in this context, all of which can be answered with the tools used above – one just has to create the appropriate 'languages'. Here is an example that gives the flavor:

> Given two sets, C and D, what is the probability of hitting 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 $\{A_i\}$, $1 \le i \le r$ be r subsets of A. What can be said about the number of trials required for the elements of at least one of them to be all obtained? One can easily write expressions as above, but direct evaluation is typically not easy.

We now turn to the minimization conjecture mentioned above:

Proof of the Conjecture

The conjecture that $E[T(\boldsymbol{p})]$ is minimal when the probability vector $\boldsymbol{p}$ is uniform is of practical interest because it would then provide an easily computable lower bound when the actual probabilities are unknown (as is the case in many applications). We mentioned the difficulties encountered in showing it from equation (1), but starting with equation (35) it is as simple as one can ask for. Denote the integrand in that equation by $f(t, \boldsymbol{p})$. We observe that $f(t, \boldsymbol{p})$ is everywhere positive, and shall show that it is minimized—uniformly in t—for the EL probability vector $\boldsymbol{e}/m$. The desired result follows. The minimization problem

$$\min_{\boldsymbol{p}\ge 0} f(t,\boldsymbol{p}) = \min_{\boldsymbol{p}\ge 0}\Big[1 - \prod_{i=1}^{m}(1-e^{-p_i t})\Big] \qquad \text{Subject to } \sum_{i=1}^{m} p_i = 1, \tag{62}$$

is clearly equivalent to the problem

$$\max_{\boldsymbol{p}\ge 0}[1 - f(t,\boldsymbol{p})] = \max_{\boldsymbol{p}\ge 0} \prod_{i=1}^{m}(1-e^{-p_i t}) \qquad \text{Subject to } \sum_{i=1}^{m} p_i = 1. \tag{63}$$

The last relation shows that at the optimum $\boldsymbol{p} > 0$. Then $f(t,\boldsymbol{p})$ is strictly positive, and since the logarithm function is strictly monotone increasing over the positive reals, we can replace the above with the equivalent problem

$$\max_{\boldsymbol{p}>0} \log[1 - f(t,\boldsymbol{p})] = \max_{\boldsymbol{p}>0} \sum_{i=1}^{m}\log(1-e^{-p_i t}) \qquad \text{Subject to } \sum_{i=1}^{m} p_i = 1. \tag{64}$$

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

$$L(\boldsymbol{p}, \lambda, \boldsymbol{\nu}) \equiv \sum_{i=1}^{m}\log(1-e^{-p_i t}) + \lambda\Big(\sum_{i=1}^{m} p_i - 1\Big) + \sum_{i=1}^{m} \nu_i p_i. \tag{65}$$

Now it is immediate to see that for any $t > 0$ the Lagrangian $L(\boldsymbol{p}, \lambda, \boldsymbol{\nu})$ has a stationary point at the uniform $\boldsymbol{p}$ (with the values for λ and $\boldsymbol{\nu}$ there uniquely defined). Moreover: in the $\boldsymbol{p}$-space it is every-where concave (its Hessian is diagonal, with negative elements only), hence that point

is a global maximum for equation (63), and a global minimum for $E[T(\boldsymbol{p})]$.

4. COMPUTATIONAL ASPECTS AND NUMERICAL EXAMPLES

A major consideration in the treatment presented above—indeed, the reason it was done in the first place—is its suitability for numerical work. We describe some computations, and remark on the significance of the numerical results.

For engineering problems that can be modelled by the CCP one needs estimates in particular for the following quantities:

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

(b) Dispersion measures of $T(\boldsymbol{p})$, 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) A rather different issue arises in applications: the vector $\boldsymbol{p}$ is frequently *unknown*. We shall consider some results that are relevant in this case.

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

The convenience of the right-hand side of equation (35) for numerical computation was already noted in [15] and was the starting point of our interest in the issue. Fig. 2 shows the integrand $f(t)$ there for the EL case with $m=100$, and the shape is typical.

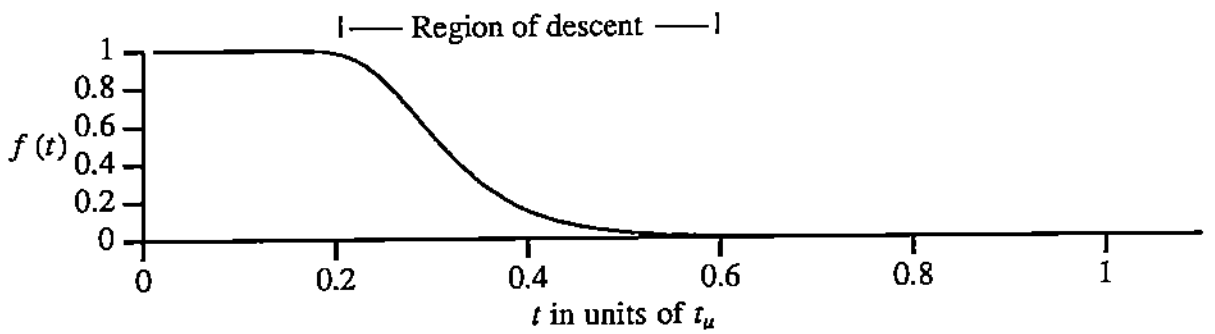

Fig. 2: The integrand of equation (35)

It 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 (for $m>1$), and vanishes exponentially. The dominant term in the tail is $\exp(-p_{\min}t)$, where $p_{\min}$ is the smallest element in $\boldsymbol{p}$. This term 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

$$\int_{t\geq t_u} e^{-p_{\min}}dt = \frac{1}{p_{\min}}e^{-t_u p_{\min}} = \varepsilon, \tag{66}$$

or $t_u=-\log(p_{\min}\varepsilon)/p_{\min}$. If there are r items associated with the minimal probability $p_{\min}$ the

cut-off point needs to be pushed to $t_u = -\log(p_{\min}\varepsilon/r)/p_{\min}$. In our work we routinely also integrated also along the interval $(t_u, 1.5t_u)$, to estimate the actual error. We used for ε the value 10^{-7}, which is far too stringent than is necessary in actual applications, but was selected in order to drive the integration procedures hard. The integration was done with the routine *quanc8* from [17], 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
$E[T(\boldsymbol{p})]$	56.24	1856.52	41,288.83	739,709.57	11,670,512.01
$mH_m\log m$	63.6	2388	51,707	901,472	13,919,295
t_u	571	17,406	281,140	4,053,471	54,629,467
# Evalu.	177	161	177	193	145

Table 1: $E[T(\boldsymbol{p})]$ 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 $mH_m\log m$, suggested in [15] as an asymptotic estimate of $E[T(\boldsymbol{p})]$. This estimate clearly tracks the correct values, but produces a substantial overestimate that seems to be little influenced by the increase of m, at least up to 10^5. The computational effort is roughly linear in m (and is essentially independent of the particular probability vector, when all the probabilities are distinct). The last line reports the number of function evaluations used in the integration, and it is practically constant in m, with the oscillations depending on the location of the region of descent of the integrand (which is fairly small compared with t_u) with respect to the evaluation points selected by the integration routine.

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.

m	10	100	1000	10,000	100,000
$E[T(\boldsymbol{p})]$	68.985	6338.74	628,226.33	62,766,148.84	6,276,050,044.96
$E[T(\boldsymbol{e}/m)]$	29.929	518.74	7485.47	97,876.06	120,901.46
t_u	1107	124,458	14,635,349	1.692×10^9	1.923×10^{11}
# Evalu.	177	225	225	209	209

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

In a sense this distribution is an inversion of the picture presented by 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. The change this causes in $E[T(\boldsymbol{p})]$ is dramatic: it is now quadratic in m, and can be shown to have the asymptotic value $m(m+1)(2\pi/\sqrt{3} - 3) \approx 0.6275987m(m+1)$. This result was stated first—without proof—in [11]. The proof consists in showing that for this

particular probability distribution, the integral in equation (35) 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 [19, 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. The rest is routine integration. The asymptotic estimate agrees with the exact value to six decimal places already for $m=100$.

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 (45) to compute the expected time to draw—for $m=1000$—all but the 10 items with smallest probabilities, and obtained $E[T_{C_{990}}] = 106{,}387.30$, about one sixth of the corresponding $E[T(\boldsymbol{p})]$.

(b) Dispersion measures and tail probabilities

The variance of $T(\boldsymbol{p})$ is somewhat more expensive to calculate—using equation (57)—than the expected value, hence we only computed it for $m=1000$ for the above distributions, and evaluated its variance ratio. We found the values 0.20521 and 0.72201 for the Zipf and the Linear distributions, respectively. (For the EL case it equals $\sqrt{1637450}/7485. = 0.17134$.) It is rarely the case that tail probabilities are easier to compute than moments, and the situation here, despite the innocent appearance of equation (55), is no exception. A direct approach is to use the Cauchy integral formula, and write

$$\begin{aligned} \text{Prob}[T(\boldsymbol{p}) > n] &= 1 - \frac{n!}{2\pi i} \oint_C z^{-n-1} \prod_{i=1}^{m} (e^{p_i z} - 1) dz. \\ &= 1 - \frac{n!}{2\pi} \int_0^{2\pi} z^{-n} \prod_{i=1}^{m} (e^{p_i z} - 1) d\theta, \qquad z = e^{i\theta}. \end{aligned} \tag{67}$$

The integration, however, except for small (and relatively uninteresting) values of m and n is numerically unstable, because the integrand oscillates, assuming very large positive and negative values along the path. The larger values occur however only at values of z with small argument, hence the integral is a good candidate for estimation by the saddle-point method: writing the integrand in the first line of equation (67) as $\exp(h_n(z))$, we have

$$\text{Prob}[T(\boldsymbol{p}) > n] = 1 - n! \frac{\prod_{i=1}^{m} (e^{p_i R_n} - 1)}{R_n^{n+1} \sqrt{2\pi h_n''(R_n)}}, \tag{68}$$

where R_n is the root of the equation $h_n'(z) = 0$. We experimented with this expression – and there were no surprises: the computations are straightforward, and reasonably accurate, but for m exceeding a few scores, and n larger than m by a single order of magnitude, one must arrange the computations in equation (68) very carefully, to avoid a numerical disaster. An example is the EL case, where it is very easy to show that $\Delta \equiv n+1-R_n > 0$ tends to zero rather fast (it is smaller than 1 already for $n \approx m\log m$, that is – for all values at which one would consider looking for tail probabilities). We find there

$$\text{Prob}[T(\boldsymbol{p}) > n] = 1 - \frac{n!R_n^{m-n}}{\Delta^m\sqrt{2\pi(n+1)(1-\Delta/m)}}. \tag{69}$$

Solving for R_n is easy, but getting reasonable accuracy for tail probabilities below 0.01 called for multiprecision arithmetic.

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.* have shown in [24] 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 $C\cdot\max(m^{-c_1/c_2}, m^{-1/2}\log m$, for some constant C.

(c) Sampling Tradeoffs

Consider equation (30). We can get closed-form results from it mainly in the EL case, for the following derivation. When $n = E[T(e/m)] = mH_m$ we find, for large m

$$E[Y_n(e/m, 1)] = m[1 - (1 - \frac{1}{m})^{mH_m}] \approx m(1 - e^{-H_m}) \tag{70}$$

The approximation of H_m by $\log m + \gamma$ (Euler's constant) will suffice here; the error is $(12m)^{-1} + O(m^{-2})$, and $\gamma = 0.57722..$. We find that the expected number of items detected by the time the CC would expect to finish is *extremely* close to m, at $m - e^{-\gamma} \approx m - 0.56146..$, with the shortfall essentially independent of m. The standard deviation of $T(e/m)$ is by equation (2) approximately equal to $m\pi/\sqrt{6} \approx 1.28255m$. It is a suitable unit of comparison with $E[T(e/m)]$, and so we compute the expected number of items found in $E[T(e/m)] + k\times m\pi/\sqrt{6}$ trials. We present the expected shortfalls below. When the expected number of detected items is $m - a$, then a is the shortfall, given in Table 3:

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 $E[T(e/m)]$ 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 five standard deviations (approximately 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 reference distributions we do not have such closed expressions, but experimentation revealed very similar patterns, usually—and surprisingly—with smaller shortfalls.

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

In many applications that are modelled by the CCP, the vector $\boldsymbol{p}$ is not known to the collector. This raises several questions of interest.

What does the item-drawing process tell us about $\boldsymbol{p}$? Of course, one 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 an inordinately long sampling time, typically much longer than $E[T(\boldsymbol{p})]$ (see [21]). We could settle for less, and just inquire about the general shape of the vector: is it close to e/m? Or to the Zipf distribution? Or to any other attracting distribution? The process $N_1(n)$ that was considered in problem (10) 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 $\boldsymbol{p}$, for a few values of m and plotted the results, in Fig. 3.

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 curious, as the others do not exhibit such a phenomenon. At least between these families one might distinguish as the sampling process continues.

Another situation of interest arises when some of the components of $\boldsymbol{p}$ might be zero! A good example of such situation is the redundancy problem introduced as application (1) in Section 1. Out of the m constraints, let q be hard ones. 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 [28] Robbins makes the following remarkable suggestion, attributed there to an observation by 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 $U(\boldsymbol{p},n)=\Sigma p_i\chi_i(n)$ – a random variable that is unobservable, by definition. (Its reciprocal is sometimes called the *resistance* of the process.)

However, let us compute its expected value: $\text{Prob}(\chi_i(n)=1)$ is simply $(1-p_i)^n$, hence

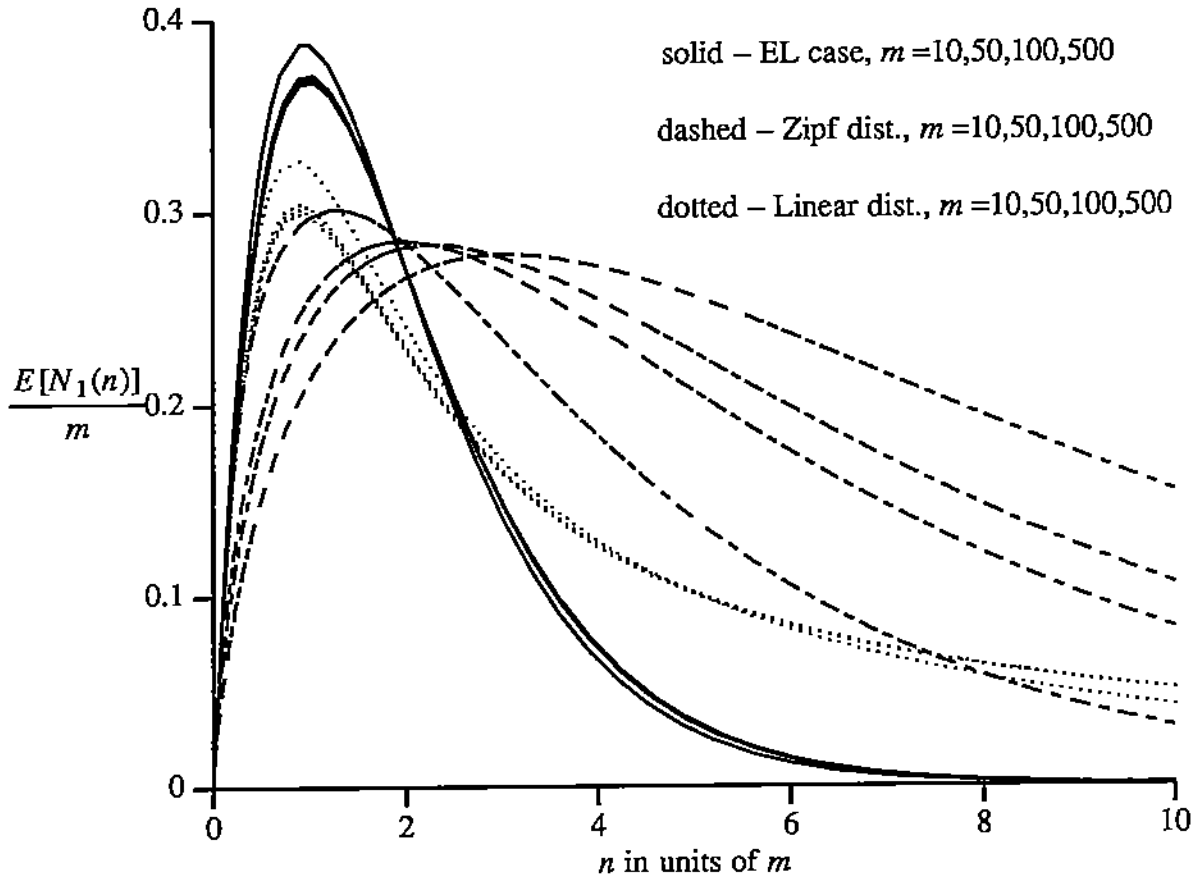

Fig. 3: The fraction of items observed exactly once .vs sample size

$$E[U(\boldsymbol{p},n)] = \sum_{i=1}^{m} p_i(1-p_i)^n. \tag{71}$$

Comparing this with equation (61) we find

$$E[U(\boldsymbol{p},n)] = \frac{1}{n+1}E[N_1(\boldsymbol{p},n+1)]. \tag{72}$$

Now, $N_1(\boldsymbol{p},n+1)$ is certainly observable, from which the optimizer has an unbiased estimate of $U(\boldsymbol{p},n)$. The curves in Fig. 3 should be viewed again in the light of this characterization. We can do even better, by considering the variance of $U(\boldsymbol{p},n)$:

$$E[U^2(\boldsymbol{p},n)] = E[(\sum_i p_i\chi_i(n))(\sum_j p_j\chi_j(n))] = \sum_{i=1}^{m} p_i^2(1-p_i)^n + \sum_{i\neq j} p_ip_j(1-p_i-p_j)^n. \tag{73}$$

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

$$E[N_2(\boldsymbol{p},n)] = \binom{n}{2}\sum_{i=1}^{m} p_i^2(1-p_i)^{n-2}$$
$$E[N_{(2)}(\boldsymbol{p},n)] = \binom{n}{2}\sum_{i\neq j} p_i p_j(1-p_i-p_j)^{n-2}. \tag{74}$$

Hence an unbiased estimate of $V[U(\boldsymbol{p},n)]$ is given by $\binom{n+2}{2}^{-2}\left(N_2(\boldsymbol{p},n+2)+N_{(2)}(\boldsymbol{p},n+2)\right)$ $-(n+1)^{-2}N_1^2(\boldsymbol{p},n+1)$.

Of even more interest is the evaluation and estimate of the variance of of the difference $U(\boldsymbol{p},n) - N_1(\boldsymbol{p},n+1)/(n+1)$. This is available from results above, except the expected value of their product, which is easy to obtain (and to estimate via equation (74)) as well:

$$E\left[U(\boldsymbol{p},n)N_1(\boldsymbol{p},n+1)\right] = \sum_{i=1}^{m} p_i^2(1-p_i)^n + \sum_{i\neq j}(1-p_i-p_j)^{n-1}\left((n+1)p_j(1-p_j) - p_i p_j\right). \tag{75}$$

Surprisingly, it is even possible to estimate the number of *new* items the sampler may expect to find in the next n' trials, once it has done n of those and dutifully recorded all of N_k. Good and Toulmin show in [18] that the suitable estimator is given by

$$\hat{N}(n') = N_1\delta - N_2\delta^2 + N_3\delta^3 \ -+ \cdots \qquad \delta \equiv \frac{n'}{n}. \tag{76}$$

This is typically useful at a stage of the sampling process when the first few N_k have not yet decreased too much towards their ultimate value – zero. This result is used in [12] to estimate the number of words of English Shakespeare knew, on the basis of his available output...

5. CONCLUSION

We have shown that for the CCP one may barter time for precision in obtaining numerical results. In some cases this was relatively easy, but some of the expressions we derived are very ill-conditioned; they may of course be handled by any computational method that uses user-defined precision, but then we are likely to run into substantial computation times again. Most of the expressions share however the fortunate characteristic, that it is precisely those large values that obstruct direct numerical evaluation, that would make them susceptible for asymptotic analysis. The way equation (68) was obtained is one example, and we expect to extend this approach to several more of the results that were derived above.

There is a different, intriguing point of view of the sampling process, that is suggested by equation (35). Consider m independent Poisson counting processes, with parameters $\{p_i\}$. Since the probabilities sum to one, the expected number of counts per time unit is one as well; let us then 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. The probability that each of the Poisson processes produces at least one count by time t is $\prod_{i=1}^{m}(1-e^{-p_i t})$. Hence the expected time until they all count (at least once) is given by equation (35). Similarly, equation (37) is the expected time to get $m-1$ distinct counts. The asymptotic distribution for the number of unsampled coupons, shown at the end of section 4(b), is

obtained by this coupling as well.

Holst shows in [23] a deeper relationship between these two schemes, and obtains a result which is related to Problem (1) above. We shall use his notation. Let $\{Z_i\}$ denote the time intervals between successive counts, and I_n – the type of the nth arrival. We consider the stopping time $T_{k;m}$, which is when exactly k of the processes have reached—or exceeded—their quota (the process of type i has a quota of r_i). Also, we let T_i denote the time until process i fills its quota, and clearly T_i has the Erlang distribution with parameters (r_i, p_i). Now he observes that $T_{k;m}$ is simply 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

$$T_{k;m} = \sum_{j=1}^{W_{k;m}} Z_j, \tag{77}$$

where $W_{k;m}$ and the Z_j are all independent. Now Holst finds for the egf of both sides

$$E\left[\exp(x\,T_{k;m})\right] = E_W\Big[E_Z\big[\exp\big(x\sum_{j=1}^{W_{k;m}} Z_j\big)\,\big|\,W_{k;m}\big]\Big], \tag{78}$$

and since the interarrival periods are iid – exp(1), this gives

$$E\left[\exp(x\,T_{k;m})\right] = E\left[(1-x)^{-W_{k;m}}\right], \tag{79}$$

from which there is an immediate relationship between the moments of $T_{k;m}$, which are straightforward to compute (in principle, that is) and the (ascending) factorial moments of $W_{k;m}$.

There is a different way to relate the two processes, which uses the Poisson transform [20]. Let $A(t)$ be a functional over the Poisson processes up to time t, such as a moment of some 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 $A(t)$ can be computed by conditioning on the number of 'arrivals' during t, since given that there occurred n arrivals, they are distributed among the coupon types according to the same underlying multinomial distribution. Hence

$$A(t) = \sum_{n\ge 0} \text{Prob}(n \text{ arrivals during } t) A_n = \sum_{n\ge 0} e^{-t}\frac{t^n}{n!}A_n. \tag{80}$$

Hence $A(t)e^t$ is the egf of $\{A_n\}$. If we can compute the first – the second is immediately available:

$$A_n = n!\,[t^n]A(t)e^t. \tag{81}$$

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 dependence), which provides a handle on the original, less tractable one. We hope to show in a forthcoming work how this may lead to further reductions in the computational effort required for some of the problems above, as well as for more complex quantities that we have not tackled so far. It is obvious however, that in terms of computational complexity this is not

a panacea; for example, it is as complicated to evaluate the moments of any order statistic of the $\{T_i\}$, when the rates (probabilities) are distinct, as it is to evaluate the right-hand side of equation (33). Hence it will *not* reduce the time required to compute the 'sample duration curve' discussed in the Remark following equation (37).

ACKNOWLEDGMENT

We are grateful to David Aldous for drawing our attention to reference [23].

REFERENCES

[1] D. Aldous, private communication, 1989.

[2] P.H. Bardell, W.H. McAnney, J. Savir: *Built-in test for VLSI – pseudorandom techniques,* J. Wiley, 1987.

[3] P.J. Bickel, J.A. Yahav: On Estimating the Number of Unseen Species – How Many Executions Were There? Technical Report No. 85 Dept. of Statistics, UC Berkeley California June 1985.

[4] A. Boneh: "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.

[5] A. Boneh: One Hit-Point Analysis (Private Communication, November 1986).

[6] A. Boneh: Prediction of the Fault-Detection Curve in Combinatorial Circuits. IBM Israel Technical Report 88.253 September 1988.

[7] R.K. Brayton: On the Asymptotic Behavior of the Number of Trials Necessary to Complete a Set with Random Selection. *Jour. Math. Anal. Appl.* **7**, 31–61 (1963)

[8] R.J. Caron, M. Hlynka, J.F. McDonald: On the Best-Case Performance of Probabilistic Methods for Detecting Necessary Constraints. Windsor Mathematics Report WMR-88-02, Dept. of Mathematics and Statistics, University of Windsor, Ontario, Canada, February 1988.

[9] R.J. Caron, J.F. McDonald: A New Approach to the Analysis of Random Methods for Detecting Necessary Linear Inequality Constraints. *Math. Prog.*, **43**, 97–102 (1989).

[10] L. Comtet: *Advanced Combinatorics*, D. Reidel, Dordrecht, 1974.

[11] F.N. David, D.E. Barton: *Combinatorial Chance,* Charles Griffin & Co. London, 1962.

[12] B. Efron, R. Thisted: Estimating the Number of Unseen Species: How Many Words Did Shakespeare Know? *Biometrika,* **63**, 435–447 ,(1976).

[13] R. D. Eldred: Test Routines Based on Symbolic Logical Statements. *Jour. Assoc. Comput.* **6**, #3, 33-66 (1959).

[14] W. Feller: *An Introduction to Probability Theory and its Applications,* Vol. 1, 3rd Ed. J. Wiley, 1968.

[15] Ph. Flajolet, D. Gardy, L. Thimonier: Birthday Paradox, Coupon Collectors, Caching Algorithms and Self-Organizing Search. INRIA RC #720. August 1987.

[16] L. Flatto: Limit Theorems for Some Random Variables Associated with Urn Models. *Ann. of Probab.* **10,** 927–934 (1982).

[17] George E. Forsythe, Michael A. Malcolm, Cleve, B. Moller: *Computer Methods for Mathematical Computations*. Prentice-Hall 1977.

[18] I.J. Good, G.H. Toulmin: The number of New Species, and the Increase in Population when a Sample is Increased. *Biometrika* **43**, 45–63 (1956).

[19] P. Henrici: *Applied and Computational Complex Analysis,* Vol. 2, J. Wiley & Sons, 1977.

[20] M. Hofri: *Probabilistic Analysis of Algorithms: On Computing Methodologies for Computer Algorithms Performance Evaluation.* Springer-Verlag, New York 1987.

[21] M. Hofri, H. Shachnai: Self-Organizing Lists and Independent References — A Statistical Synergy. The Department of Computer Science, the Technion TR#524, October 1988.

[22] L. Holst: A Unified Approach to Limit Theorems for Urn Models. *J. Appl. Probab.* **16,** 154–162, (1979).

[23] L. Holst: On Birthday, Collectors', Occupancy and Other Classical Urn Problems. *Intern. Stat. Rev.* **54**, 15–27, (1986).

[24] L. Holst, J.E. Kennedy, M.P. Quine: Rates of Convergence for Some Coverage and Urn Problems Using Coupling. *J. Appl. Probab.* **25**, 717–724, (1988).

[25] N.L. Johnson, S. Kotz: *Urn Models and their Applications. An Approach to Modern Discrete Probability Theory.* John Wiley, New York, 1977.

[26] D.J. Newman, L. Shepp: The Double Dixie Cup Problem. *Amer. Math. Month.*, **67**, 58–61 (1960).

[27] A. Rényi: Three New Proofs and a Generalization of a Theorem of Irving Weiss. *Magy. Tudo. Akad.*, **7**, 203–213 (1962).

[28] H.E. Robbins: Estimating the Total Probability of the Unobserved Outcome of an Experiment. *Ann. Math. Stat.* **39**, 256–257, (1968).

[29] R.L. Smith, J. Telgen: Random Methods for Identifying Nonredundant Constraints. Technical Report 81-4, Department of IOE, University of Michigan, Ann Arbor 1981.

[30] M. Sobel, V.R.R. Uppuluri, K. Frankowski: *Dirichlet Integrals of Type 2 and Their Applications,* Vol. IX in the series *Selected Tables in Mathematical Statistics,* Amer. Math. Soc. 1985.