Probability in the Engineering and Informational Sciences, 8, 1994, 153–168. Printed in the U.S.A.

ANALYSIS OF AN ALGORITHM FOR RANDOM SAMPLING*

CATHERINE B. HURLEY AND HOSAM M. MAHMOUD†
The George Washington University
Washington, D.C. 20052

We analyze a standard algorithm for sampling m items without replacement from a computer file of n records. The algorithm repeatedly selects a record at random from the file, rejecting records that have previously been selected, until m records are obtained. The running time of the algorithm has two components: a rejection component and a search component. We show that the probability distribution of the rejection component undergoes an infinite series of phase transitions, depending on the order of magnitude of m relative to n. We identify an infinite number of ranges of m, each with a different behavior. The rejection component is distributed as a linear combination of Poisson-like random variables. The search component is customarily done using a hash table with separate chaining. The analysis of the hashing scheme in this problem differs from previous hashing analyses, as the number of lookups in the hash table for each insertion has a geometric distribution. We show that the average overall cost of searching is asymptotically linear with only two phase transitions in the coefficient of linearity.

1. INTRODUCTION

Sampling balls without replacement from an urn is a simple task accomplished by removing balls chosen at random from the urn. Sampling records without replacement from a computer file is a more intricate task, as actual removal of

*In memory of Dr. Zaky Naguib Mahmoud.

†This author's research has been supported in part by NSA grant MDA904-92-H-3086.

records from the file is usually not allowed. Thus, the task is compounded by leaving the chosen records behind. Specifically, we wish to sample m records without replacement from a file of n records, such that all possible $\binom{n}{m}$ samples are equally likely. There are a number of standard algorithms for this problem, as surveyed by Devroye [2]. We will analyze the behavior of TRYAGAIN, which is one such algorithm.

TRYAGAIN is presented in Figure 1. Here, m items are selected without replacement from a computer file of n records, by repeatedly selecting a record at random from the file, rejecting records that have previously been selected, and trying until a new sample element is generated. This whole process is repeated until m records are obtained. Perversely, the algorithm uses sampling with replacement but in conjunction with rejection to construct a sample without replacement. Even though TRYAGAIN may use more than m selections to build the sample, its performance, as we shall demonstrate, is nevertheless competitive. The algorithm is particularly appropriate when it is not permitted to move records within the file.

For $n/2 \le m \le n$, it is more efficient to generate the "complement set," that is, use the algorithm to generate those elements that *are not in the sample*. Thus, we assume throughout that $1 \le m \le n/2$.

The algorithm has been discussed by Goodman and Hedetniemi [6] and Ernvall and Nevalainen [4]. Jakobsson [8] gave a partial average-case analysis and showed that the number of selections is linear with respect to m with a proportionality factor less than $2 \ln 2 \approx 1.38$.

After each invocation of the random number generator, TRYAGAIN has to decide whether or not the generated element is already in the sample. This membership test may be done using a bit-vector of n components with the ith entry being 1 if i is a sample element (or an index to it) or 0 otherwise. This extra space is too large if n is very large. Instead, it is preferable to cut down on this extra space by collecting only the indexes of previously generated elements in a hash table. Goodman and Hedetniemi [6] and Jakobsson [8] have suggested

```
while sample size < $m$ do
    begin
        generate $I$ uniformly from the integers 1 to $n$;
        if $I$ is not in the sample then
            add $I$ to the sample
    end;
```

FIGURE 1. TRYAGAIN—an algorithm for sampling without replacement.

the use of hashing with separate chaining with m chains. This may be particularly economical for small sample sizes.

In this paper we study the probabilistic behavior of both components of the algorithm: the overall cost of the rejection component and the overall cost of hashing. For the rejection component, we study the probability distribution of X_{nm}, the number of times the random number generator must be invoked to collect the sample; here, the number of rejections is $X_{nm} - m$. For the cost of hashing, we analyze the average behavior of Z_{nm}, the total number of comparisons made within the hash table.

We will show that the asymptotic distribution of X_{nm} undergoes an infinite series of phase transitions, depending on the relative orders of magnitude of m and n. The behavior of X_{nm} is almost ideal when $m = o(\sqrt{n})$, because the number of rejections $X_{nm} - m$ tends in probability to zero as $n \to \infty$, with convergence in rth mean holding for fixed m, for all $r > 0$. For larger samples X_{nm} departs from optimality but still remains rather good. Specifically, when m is both $\Omega(n^{(j-1)/j})$[1] and $o(n^{j/(j+1)})$ for some integer $j \geq 1$, we show that the number of rejections is approximately distributed as a linear combination of $j - 1$ Poisson random variables.

The rejection aspect of TRYAGAIN has appeared previously in the guise of the well-known coupon collectors' problem. As we shall see shortly, X_{nm} can be modeled as a sum of independent geometrically distributed random variables. Baum and Billingsley [1] analyzed this sum and obtained different limit distributions in four ranges of m. In fact, the problem has a longer history and goes back to attacks by Erdös and Rényi [3], who obtained the result of Baum and Billingsley [1] in only one range, and to attacks by Rényi [9], who found some facts about the range of asymptotic normality. Holst [7] surveyed several results on coupon collectors' problems and their connection to a number of other classical problems in probability theory. With the proper normalization within each of the four ranges, Baum and Billingsley obtained in-probability convergence for $m = o(\sqrt{n})$, convergence to a Poisson limit if $m \sim \lambda\sqrt{n}$ (for constant $\lambda > 0$),[2] and convergence in distribution to a normal limit in a range of m extending from $\Omega(n^{1/2+\epsilon})$ up to $n - o(n)$, beyond which a χ^2 limit takes over. (The range $m > n/2$ is irrelevant to our problem.) Thus, their result lumps together an infinite number of hidden phases in the range of asymptotic normality. Our alternative representation of X_{nm} as an infinite sum of independent Poisson-like random variables refines this range of asymptotic normality.

As for the cost of hashing, we shall take a look at the average of Z_{nm}. It will turn out that $\mathrm{E}[Z_{nm}]/m \to c$, a constant, and the constant c itself has two phase transitions: it is asymptotic to $(1 - 1/m)/2$ if m is fixed, then becomes asymptotic to $\frac{1}{2}$ if m goes to infinity but $m = o(n)$, and then becomes a function of λ when $m \sim \lambda n$.

The paper will be organized as follows. In Section 2 we establish our notation. In Section 3 the behavior of X_{nm} is given in the exact form as the infinite sum of Poisson-like random variables. The result is easy to apply for small sample sizes, as we illustrate when m is both $\Omega(n^{(j-1)/j})$ and $o(n^{j/(j+1)})$ for $j = 1,2,3$. For larger samples the normal limiting distribution provides an easier approximation, and this result is given in Section 4. In Section 5 we analyze the average cost of hashing. We will show there that the cost of hashing is asymptotically linear in m with an explicit characterization of the coefficient of linearity and the phase transitions therein. Section 6 is a discussion about the practical significance of the results.

2. THE TECHNICAL SETUP

Throughout the paper we shall use the following standard notation. We shall denote the harmonic numbers $\sum_{j=1}^{k} 1/j$ by H_k, and $H_k^{(2)}$ will denote $\sum_{j=1}^{k} 1/j^2$, the second order harmonic numbers. Geometric(p) will denote a geometrically distributed random variable with rate of success p per trial; Poisson(λ) will denote a Poisson random variable with parameter λ; the symbol $N(0,1)$ will denote a random variable having the standard normal distribution; and Hypergeometric(N, n, k) will denote a random variable with a hypergeometric distribution, that is, the number of type 1 members in a subset of size k taken from a population of size N and having n members of type 1, with the rest of type 2. Convergence in distribution and in probability will be denoted by the symbols $\xrightarrow{\mathfrak{D}}$ and $\xrightarrow{\mathcal{P}}$, as usual. On the other hand, the symbol $\overset{\mathfrak{D}}{=}$ will denote exact equality in distribution.

When TRYAGAIN is in its ith stage, that is, after it has picked $i-1$ sample elements and is about to choose the ith sample element, it repeatedly generates a random number until a new sample element is obtained. Let G_i be the number of iterations required for the ith sample element. Clearly,

$$G_i \overset{\mathfrak{D}}{=} \text{Geometric}\left(\frac{n-i+1}{n}\right).$$

The random variables G_i, $i = 1, \ldots, m$, are independent, and X_{nm}, the total number of generations, is

$$X_{nm} = G_1 + G_2 + \cdots + G_m. \tag{1}$$

As in Feller [5], the mean and variance can be written in terms of harmonic numbers as

$$\begin{aligned} \mathbf{E}[X_{nm}] &= n(H_n - H_{n-m}), \\ \mathbf{Var}[X_{nm}] &= n^2(H_n^{(2)} - H_{n-m}^{(2)}) - n(H_n - H_{n-m}). \end{aligned}$$

3. THE COST OF REJECTION FOR SMALL SAMPLES

Let $\psi_{nm}(t)$ be the moment generating function of $X_{nm} - m$. From the representation of X_{nm} in Eq. (1),

$$\begin{aligned}\psi_{nm}(t) &= e^{-mt}\mathbf{E}[e^{(G_1+G_2+\cdots+G_m)t}]\\ &= e^{-mt}\prod_{k=1}^{m}\mathbf{E}[e^{G_k t}]\\ &= \prod_{k=1}^{m}\frac{n-k+1}{n-(k-1)e^t}, \quad \text{for } t<\ln 2,\end{aligned}$$

using the fact that a Geometric(p) has the moment generating function $pe^t/(1-(1-p)e^t)$. We shall next examine the behavior of the product in the last quantity.

Lemma 1: *For any fixed $t<\ln 2$, we have*

$$\prod_{k=1}^{m}\frac{n-k+1}{n-(k-1)e^t} = \exp\left\{\sum_{k=1}^{m}\sum_{j=1}^{\infty}\left(\frac{k-1}{n}\right)^j\frac{e^{jt}-1}{j}\right\}.$$

Proof: Let

$$S_{nm}(t) = \prod_{k=1}^{m}[n-(k-1)e^t].$$

Taking logarithms, one finds

$$\begin{aligned}\ln S_{nm}(t) &= \sum_{k=1}^{m}\ln\left[n\left(1-\frac{k-1}{n}e^t\right)\right]\\ &= m\ln n + \sum_{k=1}^{m}\ln\left(1-\frac{k-1}{n}e^t\right)\\ &= m\ln n - \sum_{k=1}^{m}\sum_{j=1}^{\infty}\left(\frac{k-1}{n}\right)^j\frac{e^{jt}}{j}.\end{aligned} \tag{2}$$

The lemma now follows from calculating $S_{nm}(0)/S_{nm}(t)$. ■

So we have

$$\psi_{nm}(t) = \exp\left\{\sum_{j=1}^{\infty}\theta_{nm}(j)(e^{jt}-1)\right\},$$

where

$$\theta_{nm}(j) \overset{\text{def}}{=} \frac{1}{j}\sum_{k=1}^{m}\left(\frac{k-1}{n}\right)^j,$$

but $\exp\{\lambda(e^{jt}-1)\}$ is the moment generating function of j times a Poisson(λ) random variable. Thus, the number of rejections has a moment generating function coinciding with that of a linear combination of infinitely many independent Poisson random variables. We have thus proved the following theorem.

Theorem 1: *The number of rejections is distributed as a linear combination of infinitely many independent Poisson random variables*:

$$X_{nm} - m \overset{\mathfrak{D}}{=} \sum_{j=1}^{\infty} j\,\mathrm{Poisson}\left(\frac{1}{j}\sum_{k=1}^{m}\left(\frac{k-1}{n}\right)^{j}\right).$$

We now apply Theorem 1 to the small sample situation, which we define as $m = o(n)$. The interpretation of Theorem 1 depends on the range of m. Roughly speaking, for any fixed j, the random variables $j\mathrm{Poisson}(j^{-1}\times \sum_{k=1}^{m}((k-1)/n)^{j}) = j\mathrm{Poisson}(O(m^{j+1}/n^{j}))$ are "very small" with high probability if $m = o(n^{j/(j+1)})$. Thus, if m is both $\Omega(n^{(j-1)/j})$ and $o(n^{j/(j+1)})$ for some positive integer j, then the first $j - 1$ Poisson random variables dominate and the others are asymptotically negligible. For each value of j, the orders of magnitude between $\Omega(n^{(j-1)/j})$ and $o(n^{j/(j+1)})$ correspond to a new range (the jth range) of m, with $\Omega(n^{(j-1)/j})$ as its lower boundary. Every time m "crosses" the lower boundary into a new range, one extra random variable, which is a multiple of a Poisson random variable, is "released" into the picture. This curious behavior is illustrated later by a few corollaries concerning the behavior in the first few ranges.

We now provide a rigorous justification for neglecting the tail terms in the series of Poisson random variables of Theorem 1. The following asymptotic representation of $\psi_{nm}(t)$ is helpful. The negative of the logarithm in Eq. (2) may be written as

$$-\ln\left(1-\frac{k-1}{n}e^{t}\right) = \sum_{j=1}^{r-1}\left(\frac{k-1}{n}\right)^{j}\frac{e^{jt}}{j} + O\left(\left(\frac{k}{n}\right)^{r}\right)$$

and

$$-\sum_{k=1}^{m}\ln\left(1-\frac{k-1}{n}e^{t}\right) = \sum_{j=1}^{r-1}\sum_{k=1}^{m}\left(\frac{k-1}{n}\right)^{j}\frac{e^{jt}}{j} + O\left(\frac{m^{r+1}}{n^{r}}\right).$$

Thus,

$$\psi_{nm}(t) = \exp\left\{\sum_{j=1}^{r-1}\sum_{k=1}^{m}\left(\frac{k-1}{n}\right)^{j}\frac{e^{jt}-1}{j} + O\left(\frac{m^{r+1}}{n^{r}}\right)\right\}. \qquad (3)$$

In the special case of a very small sample, we now have the following corollary.

Corollary 1: *If $m = o(\sqrt{n})$, then the number of rejections $X_{nm} - m \xrightarrow{\mathcal{P}} 0$.*

Proof: In Eq. (3) take $r = 1$ to obtain

$$\psi_{nm}(t) = \exp\left\{O\left(\frac{m^{2}}{n}\right)\right\} \sim 1.$$

Hence, $X_{nm} - m \xrightarrow{\mathfrak{D}} 0$ and, as convergence in distribution to a constant implies convergence in probability, we have $X_{nm} - m \xrightarrow{\mathcal{P}} 0$. ■

Remark: In the case of fixed m, for every $r > 0$ the uniform integrability of X^r_{nm} is demonstrated in the Appendix. This fact, together with Corollary 1, shows that for fixed m the random variable X_{nm} converges in L^r for every $r > 0$, that is, $\mathbf{E}[X^r_{nm}] \to m^r$.

Applying Eq. (3) to the next range of m values (i.e., $r = 2$), we have the following corollary.

COROLLARY 2. *If m is both $\Omega(\sqrt{n})$ and $o(n^{2/3})$, then $\psi_{nm}(t) \sim \exp\{(m^2/2n)(e^t - 1)\}$.*

As a Poisson random variable with a large parameter is approximately normally distributed, we can interpret Corollary 2 as follows:

$$X_{nm} - m \xrightarrow{\mathfrak{D}} \text{Poisson}(\lambda^2/2), \quad \text{if } m \sim \lambda\sqrt{n},$$

otherwise the Poisson parameter $m^2/2n$ tends to infinity with n, and so

$$\frac{X_{nm} - m}{\sqrt{m^2/2n}} \xrightarrow{\mathfrak{D}} N(0,1).$$

For the next higher range of m (with $r = 3$), Eq. (3) yields the next corollary.

COROLLARY 3: *If m is $\Omega(n^{2/3})$ and $o(n^{3/4})$,*

$$\psi_{nm}(t) \sim \exp\left\{\frac{m(m-1)}{2n}(e^t - 1) + \frac{m(m-1)(2m-1)}{12n^2}(e^{2t} - 1)\right\}.$$

Observe that the right-hand side in the preceding expression is the moment generating function of a combination of two independent Poisson random variables: Poisson$((m(m-1))/2n) + 2$ Poisson$((m(m-1)(2m-1))/12n^2)$. Asymptotic normality follows after suitable normalization. One may continue to obtain similar expressions in higher ranges of m.

Of course, there is only one limiting distribution in the range of m beginning at $\Omega(\sqrt{n})$, which is the Normal distribution. That Normal distribution captures the asymptotic behavior of the first Poisson random variable in the series of Theorem 1. However, the existence of lower order Poisson random variables may affect the rate of convergence and, in general, including those variables may provide a better approximation for small values of n. As an example, Table 1 compares three approximations to $P(X_{nm} \le m)$, with the exact probability (computed using the expression given in the Appendix), for a sequence of increasing n values and with $m = n^{2/3}$. In each case the two-Poisson approximation to the distribution of X_{nm} provides quite accurate results.

TABLE 1. Comparison of Three Approximations to $P(X_{nm} \leq m)$, With the Exact Probability[a]

n	Exact	Normal	One Poisson	Two Poissons
2^3	0.4102	0.1950	0.4724	0.4234
2^6	0.1290	0.0859	0.1534	0.1318
2^9	0.0164	0.0237	0.0195	0.0166
2^{12}	2.92E−4	2.38E−3	3.46E−4	2.93E−4

[a] Column 1 gives the exact probability; column 2 uses the asymptotic Normal distribution given by Baum and Billingsley [1], while columns 3 and 4 use the one- and two-term approximations, respectively, given in the result of Theorem 1. In each case $m = n^{2/3}$.

4. THE COST OF REJECTION FOR LARGE SAMPLES

Under the assumption of a large sample, that is, $m \sim \lambda n$ for some $\lambda \in (0,\frac{1}{2}]$, we will show in this section that the algorithm running time converges in distribution to a normally distributed random variable. This could perhaps be approached by obtaining a limit from Eq. (3) of the previous section, for increasingly higher ranges of m. However, here we choose to establish the result via Lindeberg's condition, a route considered prohibitive by Baum and Billingsley [1]. Formally stated, we shall prove the following theorem.

THEOREM 2: *When $m \sim \lambda n$ where $\lambda \in (0,\frac{1}{2}]$,*

$$\frac{X_{nm} - n\ln\left(\dfrac{1}{1-\lambda}\right)}{\sqrt{n\left(\dfrac{\lambda}{1-\lambda} - \ln\left(\dfrac{1}{1-\lambda}\right)\right)}} \xrightarrow{\mathfrak{D}} N(0,1).$$

PROOF: From the representation of X_{nm} in Eq. (1), it is sufficient to verify Lindeberg's condition to establish the asymptotic normality of

$$\frac{X_{nm} - \mathbf{E}[X_{nm}]}{\sqrt{\mathbf{Var}[X_{nm}]}},$$

as n and hence $m \to \infty$. As shown by Baum and Billingsley [1], the mean and variance satisfy the following asymptotic relations:

$$\mathrm{E}[X_{nm}] \sim n\ln\left(\frac{1}{1-\lambda}\right) \tag{4}$$

$$\mathbf{Var}[X_{nm}] \sim n\left(\frac{\lambda}{1-\lambda} - \ln\left(\frac{1}{1-\lambda}\right)\right) \overset{\text{def}}{=} c_\lambda n, \tag{5}$$

and asymptotic normality in the form stated will follow.

To simplify notation we introduce $\sigma_{nm}^2 = \mathbf{Var}[X_{nm}]$ and $\mu_k = \mathbf{E}[G_k]$. Fix $\epsilon > 0$, and define "Lindeberg's quantity" $L_{nm}(\epsilon)$ as

$$L_{nm}(\epsilon) = \frac{1}{\sigma_{nm}^2} \sum_{k=1}^{m} \sum_{|x-\mu_k| \geq \epsilon\sigma_{nm}} (x-\mu_k)^2 P(G_k = x).$$

We now verify that

$$\lim_{m\to\infty} L_{nm}(\epsilon) = 0.$$

For n sufficiently large, $\{x : x \leq \mu_k - \epsilon\sigma_{nm}\}$ is a set of negative integers for any k with $1 \leq k \leq m$ and $P(G_k = x)$ for any x from that set is 0—this holds for large n because

$$\mu_k = \frac{n}{n-k+1} < 2 < \epsilon\sigma_{nm},$$

and, according to Eq. (5), $\sigma_{nm} \to \infty$ as $n \to \infty$. Therefore, for large n

$$\begin{aligned} L_{nm}(\epsilon) &= \frac{1}{\sigma_{nm}^2} \sum_{k=1}^{m} \sum_{x \geq \mu_k + \epsilon\sigma_{nm}} (x-\mu_k)^2 P(G_k = x) \\ &= \frac{1}{\sigma_{nm}^2} \sum_{k=1}^{m} \sum_{y \in Y_k(n,m)} y^2 p_k q_k^{y+\mu_k-1}, \end{aligned}$$

where

$$Y_k(n,m) = \{y = j - \mu_k : j \text{ is a positive integer and } j \geq \mu_k + \epsilon\sigma_{nm}\},$$

$p_k = (n-k+1)/n$ and $q_k = 1 - p_k$. Thus,

$$L_{nm}(\epsilon) = \frac{1}{\sigma_{nm}^2} \sum_{k=1}^{m} p_k q_k^{\mu_k+1} \sum_{y \in Y_k(n,m)} y^2 q_k^{y-2}.$$

Using the inequality $y^2 \leq 2y(y-1)$, valid for any $y \geq 2$, and the identity

$$\begin{aligned} \sum_{y \geq y_0} y(y-1)q^{y-2} &= \frac{q^{y_0-2}}{(1-q)^3} \{(1-q)^2 y_0(y_0-1) + 2q(y_0 - qy_0 + q)\} \\ &< \frac{q^{y_0-2}}{(1-q)^3} 2y_0^2, \quad \text{for } y_0 \geq 1, \end{aligned}$$

we can write

$$L_{nm}(\epsilon) = \frac{4}{\sigma_{nm}^2} \sum_{k=1}^{m} p_k q_k^{\mu_k+1} \left[\frac{q_k^{y_k(n,m)-2}}{p_k^3} y_k^2(n,m)\right],$$

where $y_k(n,m) = \min\{y \in Y_k(n,m)\}$, which is ≥ 2 for large n. Therefore, as $\mu_k \geq 1$ for all $1 \leq k \leq m$,

$$L_{nm}(\epsilon) < \frac{4}{\sigma_{nm}^2} \sum_{k=1}^{m} \left(\frac{k-1}{n}\right)^{y_0-1} \left(\frac{n}{n-k+1}\right)^2 y_0^2,$$

where $\epsilon\sigma_{nm} \leq y_0 = \sup_k y_k(n,m) < \epsilon\sigma_{nm} + 1$. Recalling that $1 \leq m \leq n/2$, we have

$$\begin{aligned} L_{nm}(\epsilon) &< \frac{4y_0^2 n^2}{\sigma_{nm}^2 (n/2+1)^2} \sum_{k=1}^{m} \left(\frac{k}{n}\right)^{y_0-1} \\ &< \frac{16 y_0^2}{\sigma_{nm}^2 n^{y_0-1}} \sum_{k=1}^{m} k^{y_0-1} \\ &< \frac{16 y_0^2}{\sigma_{nm}^2 n^{y_0-1}}\, m \left(\frac{n}{2}\right)^{y_0-1} \\ &\leq \frac{16 y_0^2 n}{\sigma_{nm}^2 2^{y_0-1}} \\ &< \frac{16(\epsilon\sigma_{nm}+1)^2 n}{\sigma_{nm}^2 2^{\epsilon\sigma_{nm}}} \\ &< \frac{16(2\epsilon\sigma_{nm})^2 n}{\sigma_{nm}^2 2^{\epsilon\sigma_{nm}}} \\ &= \frac{64\epsilon^2 n}{2^{\epsilon\sigma_{nm}}}. \end{aligned}$$

But $\sigma_{nm} \sim \sqrt{c_\lambda n} \to \infty$ as $n \to \infty$ (cf. Eq. (5)), and so $\lim_{n\to\infty} L_{nm}(\epsilon) = 0$, for any $\epsilon > 0$. ■

5. THE COST OF HASHING

Among several possible choices for a hashing scheme, authors who worked on this problem preferred the method of hashing with separate chaining. The method is quite simple and efficient. A hash table of m slots labeled $1, \ldots, m$, is set up. The ith slot contains a header pointer to a linked list of sample elements that hash to position i. The hash function is $h(x) = \lceil xm/n \rceil$. At the kth stage, each failure in generating a new sample element corresponds to a successful search in the table. That search is handled as follows. The hash function is invoked and a sequential search in the appropriate list is begun at its header. At the end of the kth stage, a new sample element is generated and inserted in the table in the following manner. The new element is first hashed. The linked list corresponding to the hash position is searched until it is exhausted and the **NIL** pointer at its tail is reached. Then, the new element is inserted in the list. The most efficient way is to insert it at the beginning of the list, as this does not require any additional search for an insertion position.

Consider the situation at the kth stage. The random number generator is invoked G_k times, of which the first $G_k - 1$ are failures (producing an element already in the hash table) and the last time is a success (producing a new sample element). The hash table has m slots, where each slot contains a linked list of at most $\lceil n/m \rceil$ sample elements. For simplicity we assume that n is an exact multiple of m, and we define s as n/m. (If m is not an exact multiple of n, the results will still hold asymptotically.)

We wish to study the mean of Z_{nm}, the total number of comparisons used within the hash table. Writing C_k as the total number of comparisons required for inserting the kth sample element, we have

$$Z_{nm} = C_1 + C_2 + \cdots + C_m.$$

We derive expressions for the mean of C_k.

Let $h_{k,i}$, $i = 1, \ldots, G_k - 1$, be the number of hash table comparisons required for the ith rejection at the kth stage. Suppose that when the kth sample element is finally obtained it calls for $\tilde{h}_k$ comparisons. Then we have

$$C_k = h_{k,1} + h_{k,2} + \cdots + h_{k,G_k-1} + \tilde{h}_k. \tag{6}$$

Let $\mathcal{F}_{k-1}$ be the σ-field generated by the first $k-1$ stages. Note that $\mathcal{F}_{k-1}$ includes the complete history of the process up to stage $k-1$; in particular, it contains the depths of the linked lists $d_{k-1,1}, d_{k-1,2}, \ldots, d_{k-1,m}$, after the $(k-1)$st sample element is obtained. For short, we will use the notation d_1, $d_2, \ldots, d_m$, for these depths, where the dependence on k is understood.

We thus have

$$\mathbf{E}[\tilde{h}_k \mid \mathcal{F}_{k-1}] = \sum_{i=1}^{m} \frac{d_i(s-d_i)}{n-k+1},$$

because if the ith list contains d_i elements, only $s - d_i$ of the remaining $n-k+1$ equally likely unselected elements are candidates to be placed in the ith linked list. Hence,

$$\begin{aligned}\mathbf{E}[\tilde{h}_k] &= \frac{1}{n-k+1}\left(\sum_{i=1}^{m} (s\mathbf{E}[d_i] - \mathbf{E}[d_i^2])\right)\\ &= \frac{m}{n-k+1}(s\mathbf{E}[d_1] - \mathbf{E}[d_1^2]),\end{aligned}$$

because the d_i's have the same distribution.

We next obtain the expected value of the part of Eq. (6) corresponding to the comparisons associated with rejection. Conditioning, we have

$$\mathbf{E}\left[\sum_{i=1}^{G_k-1} h_{k,i} \,\middle|\, G_k - 1\right] = (G_k - 1)\mathbf{E}[h_{k,1}],$$

because the $h_{k,i}$'s have the same distribution, which is independent of $G_k - 1$. Therefore,

$$\mathbf{E}\left[\sum_{i=1}^{G_k-1} h_{k,i}\right] = \mathbf{E}[G_k - 1]\mathbf{E}[h_{k,1}] = \frac{k-1}{n-k+1}\,\mathbf{E}[h_{k,1}].$$

At the kth stage, there are $k - 1$ sampled elements of which d_i belong to the ith slot. Also, the number of comparisons required for a sampled element from the ith slot has a discrete uniform distribution on $\{1,\ldots,d_i\}$ with a mean value of $(d_i + 1)/2$. Thus,

$$\mathbf{E}[h_{k,1}\,|\,\mathcal{F}_{k-1}] = \sum_{i=1}^{m} \frac{d_i+1}{2} \times \frac{d_i}{k-1} = \frac{1}{2(k-1)} \sum_{i=1}^{m} (d_i^2 + d_i)$$

and

$$\mathbf{E}[h_{k,1}] = \frac{m}{2(k-1)}\,(\mathbf{E}[d_1^2] + \mathbf{E}[d_1]).$$

Hence,

$$\begin{aligned}\mathbf{E}[C_k] &= \frac{m}{2(n-k+1)}\,(\mathbf{E}[d_1^2] + \mathbf{E}(d_1]) + \frac{m}{n-k+1}\,(s\mathbf{E}[d_1] - \mathbf{E}[d_1^2]) \\ &= \frac{m}{n-k+1}\left(s\mathbf{E}[d_1] - \frac{1}{2}\,\mathbf{E}[d_1(d_1-1)]\right).\end{aligned}$$

The random variable d_1 follows a Hypergeometric$(n,s,k-1)$ distribution, and so $\mathbf{E}[d_1] = (k-1)s/n$ and $\mathbf{E}[d_1(d_1-1)] = (k-1)(k-2)s(s-1)/n(n-1)$. Substituting s by n/m, we obtain, after some algebra,

$$\mathbf{E}[C_k] = \frac{(k-1)(2n^2 - kn + mk - 2m)}{2(n-k+1)m(n-1)}.$$

We next show that $\mathbf{E}[C_k]$ is uniformly bounded for all relevant values of k, thus identifying a constant for the $O(1)$ bound discussed by Jakobsson [8].

Lemma 2: *For* $1 \le k \le m \le n/2$,

$$\mathbf{E}[C_k] < \tfrac{7}{4}.$$

Proof: First, write

$$\mathbf{E}[C_k] = \frac{k-1}{2} \times \left[\frac{n}{(n-k+1)m} + \frac{n}{(n-1)m} + \frac{k-2}{(n-k+1)(n-1)}\right].$$

Because $k \le m \le n/2$, and $2(n-k+1) \ge n+2$, we have

$$\mathbf{E}[C_k] \le 1 + \frac{(m-1)n}{2m(n-1)} + \frac{(m-2)(m-1)}{(n+2)(n-1)}.$$

Substituting $m = n/2$, we obtain the desired bound. ■

Although one could quickly get the linear upper bound $7m/4$ for the average overall number of comparisons, one expects that the coefficient of linearity can be improved because $\frac{7}{4}$ is a uniform bound at all stages; it is plausible that $\frac{7}{4}$ is a loose bound at the early stages. The overall average of the number of comparisons may be obtained by adding up $\mathbf{E}[C_k]$, for $k = 1,2,\ldots,m$. After somewhat long and unpleasant algebra, we obtain the next theorem.

THEOREM 3:

$$\mathbf{E}[Z_{nm}] = \frac{n(n+m)}{2m}(H_n - H_{n-m}) - \frac{2n^2 + mn - n + m^2 - 3m}{4(n-1)}.$$

Asymptotically, we have the following corollary.

COROLLARY 4:

$$\mathbf{E}[Z_{nm}] \sim \begin{cases} \dfrac{m-1}{2}, & \textit{if } m \textit{ is fixed}; \\ \dfrac{m}{2}, & \textit{if } m \to \infty, \textit{ and } m = o(n); \\ \left[\dfrac{\lambda+1}{2\lambda^2}\ln\left(\dfrac{1}{1-\lambda}\right) - \dfrac{\lambda^2+\lambda+2}{4\lambda}\right]m, & \textit{if } m \sim \lambda n. \end{cases}$$

PROOF: Using the asymptotic expansion of the harmonic numbers, and the fact that $m \le n/2$, we obtain

$$H_n - H_{n-m} = \ln\left(\frac{n}{n-m}\right) - \frac{m}{2n(n-m)} + \frac{m(2n-m)}{12n^2(n-m)^2} + O\left(\frac{1}{n^3}\right).$$

The lemma follows from straightforward algebraic manipulation using Taylor's expansion with remainder for the logarithm. ■

6. DISCUSSION

We studied the behavior of an algorithm for sampling m records from a computer file of n records. For very small samples $(m = o(\sqrt{n}))$, Corollary 1 shows that the algorithm is almost ideal; it takes m steps in probability, with convergence in all moments for fixed m. In the higher ranges of m, some variability appears and the algorithm may take more than m steps. Jakobsson [8] showed that the average number of steps is less than cm for a constant c not exceeding $2\ln 2 \approx 1.38$. In fact, relation (4) shows that the upper bound of

Jakobsson is sharp if $\lambda = \frac{1}{2}$ but can be improved for $0 \le \lambda < \frac{1}{2}$. For example, if $m = \lceil n/10 \rceil$, the algorithm takes about $1.05m$ steps on average, only 5% worse than the ideal case of finishing in m steps. In addition, it is not likely for the algorithm to deviate by large amounts from its mean. For example, if $m = \lceil n^{0.6} \rceil$, from Corollary 2 the number of rejections is asymptotically distributed as Poisson$(0.5n^{0.2})$, and, according to Chebychev's inequality, the probability the algorithm takes $\sqrt{m}$ steps beyond its average is

$$P(X_{nm} - \mathbf{E}[X_{nm}] \ge \sqrt{m}) \le \frac{n^{0.2}}{2n^{0.6}} \to 0, \quad \text{as } n \to \infty.$$

In the worst case scenario where $m = n/2$, Theorem 2 shows that, for large values of n,

$$P(X_{nm} - \mathbf{E}[X_{nm}] \ge 2\sqrt{m}) \approx .0052.$$

The distributional theory of the hashing aspect of this problem is harder to analyze. We have only been able to analyze it on average and found out that the overall cost of hashing is asymptotically linear in m. The coefficient of linearity is a constant slightly less than $\frac{1}{2}$, if m is fixed, and is $\frac{1}{2}$ if m goes to infinity but in such a way that $m = o(n)$; when $m \sim \lambda n$, the coefficient of linearity goes up a little bit. For example, $\mathbf{E}[Z_{nm}] \sim (3 \ln 2 - \frac{11}{8})m \approx .704m$, if $n = 2m$.

The real cost of the algorithm is a linear combination of X_{nm} and Z_{nm}. Thus, a complete distributional analysis will further require finding the joint distribution of X_{nm} and Z_{nm}, which appears to be a very formidable task. However, we have all the desired average-case results. The coefficients used in the linear combination depend on the speed of the particular computer in use because the operations involved in random number generation are quite different from those involved in hashing.

Acknowledgments

The authors would like to thank R. Modarres and R. T. Smythe for helpful discussions.

NOTES

1. $\Omega(g(n)) = \{f(n) : \exists c > 0,\ n_0 > 0$ such that $\forall n \ge n_0,\ 0 \le cg(n) \le f(n)\}$.
2. If $\lim_{n\to\infty} (f(n))/(g(n)) = 1$, then $f(n)$ and $g(n)$ are said to be asymptotically equivalent, denoted by $f(n) \sim g(n)$.

References

1. Baum, L. & Billingsley, P. (1965). Asymptotic distributions for the coupon collector's problem. *Annals of Statistics* 36: 1835–1839.
2. Devroye, L. (1986). *Non-uniform random variate generation*. New York: Springer-Verlag.
3. Erdös, P. & Rényi, A. (1961). On a classical problem of probability theory. *Magyar Tudomanyos Akademia Mat. Kutató Int. Közl.* 6: 215–220.

4. Ernvall, J. & Nevalainen, O. (1982). An algorithm for unbiased random sampling. *The Computer Journal* 25: 45-47.
5. Feller, W. (1968). *An introduction to probability theory and its applications*, Vol. 1, 3rd ed. New York: Wiley.
6. Goodman, S. & Hedetniemi, S. (1977). *Introduction to the design and analysis of algorithms.* New York: McGraw-Hill.
7. Holst, L. (1986). On birthday, collectors', occupancy and other classical urn problems. *International Statistical Review* 54: 15-27.
8. Jakobsson, M. (1985). Sampling without replacement in linear time. *The Computer Journal* 28: 412-413.
9. Rényi, A. (1962). Three new proofs and a generalization of a theorem of Irving Weiss. *Magyar Tudomanyos Akademia Mat. Kutató Int. Közl.* 7: 203-214.

APPENDIX

Here we will demonstrate the uniform integrability of X_{nm}^r for $0 < r < \infty$ and fixed m.

The distribution of X_{nm}, the total number of generations, is given by (see, e.g., Feller [5, pp. 57-59])

$$P(X_{nm} = k) = \binom{n}{m} \frac{m!S(k-1,m-1)}{n^k},$$

where $S(k,m)$ is a Stirling number of the second kind, that is, the number of ways of distributing k items into m boxes, leaving no empty box. Clearly, $S(k,m) < m^k/m!$, the number of ways of distributing k items into m boxes, and so

$$P(X_{nm} = k) < \binom{n}{m}\left(\frac{m}{n}\right)^k < \frac{m^m}{m!}\left(\frac{m}{n}\right)^{k-m} < \frac{m^m}{m!}\left(\frac{1}{2}\right)^{k-m},$$

because $m \leq n/2$. Then, we have that

$$\begin{aligned} \mathbf{E}[X_{nm}^r \mathbf{1}_{\{X_{nm} \geq \alpha\}}] &< \frac{m^m}{m!} \sum_{k \geq \alpha} \left(\frac{1}{2}\right)^{k-m} k^r \\ &= \frac{m^m}{m!} \sum_{k \geq \alpha - m} \left(\frac{1}{2}\right)^k (k+m)^r. \end{aligned}$$

Because the series $\sum_k (\frac{1}{2})^k (k+m)^r$ is convergent, we obtain

$$\lim_{\alpha \to \infty} \sup_n \mathbf{E}[X_{nm}^r \mathbf{1}_{\{X_{nm} \geq \alpha\}}] = 0. \qquad \blacksquare$$