Applied Probability Trust (5 January 2010)

QUOTA FULFILLMENT, RANDOM WALKS AND MATCHING SOCKS

JAMES J. MADDEN,* *Louisiana State University*

Abstract

Suppose an urn contains n sets of beads, each set of size k. Beads are drawn from the urn one at a time without replacement. We derive a closed-form expression for the expected number of draws required to completely remove the first k-set, the second, the third, etc. We accomplish this by recasting the problem as a special kind of random walk through the integer points in $[0,k]^n$ and exploiting the symmetries of the hypercube. We also show that the expected proportion of beads removed when the first k-set is completed is between two constant multiples of $\sqrt[k]{1/n}$.

Keywords: quota fulfillment; hypergeometric distribution; urn model

2000 Mathematics Subject Classification: Primary 60C05
Secondary 60G50

1. Introduction

Thiago Hirai, a young engineer in Seattle, writes at his blog:

> I have the habit of leaving my clean clothes in the dryer until they run out or until I need to dry some more ... I rarely spend the time to put them away. When I need some socks I go to the dryer and try to find a matching pair. This process is often more difficult than I'd expect and it has bothered me for years. So today I decided to compute the probabilities involved and the expected number of draws until I find a matching pair. [2]

Mr. Hirai proceeds to derive a recursive formula similar to equation (4) below. This problem is a kind of *quota fulfillment problem*. Such problems are considered in some generality in [1], but I have not found an explicit formula for the expectation that Mr. Hirai seeks in any printed source.

In the present paper, we consider a more general quota fulfillment problem. Suppose an urn contains n sets of beads, each set of size k. Beads are drawn from the urn one at a time without replacement. At any stage each remaining bead has an equal chance of being drawn. Let $e(n,k,c)$ denote the expected number of draws required to remove c entire k-sets. The socks problem asks for $e(n,2,1)$. We shall prove that if $n \geq c \geq 1$ and $k \geq 1$:

$$e(n,k,c) \quad = \quad \frac{c\,k}{1} \cdot \frac{(c+1)\,k}{1+c\,k} \cdots \frac{(n-1)\,k}{1+(n-2)\,k} \cdot \frac{n\,k}{1+(n-1)\,k}. \tag{1}$$

In order to do this, we recast the problem as an "ascending random walk" through the set of lattice points in an n-dimensional box. This provides a possible approach to

* Postal address: Department of Mathematics, Louisiana State University, Baton Rouge LA, 70803, USA

other quota fulfillment problems, such as those that arise when the sets of beads all have different sizes. The technique generalizes to ascending walks in any finite poset and raises several interesting questions that will be addressed elsewhere. I would like to thank Ambar Sengupta for suggesting this idea.

2. The socks problem

In this section, we present a self-contained solution of the socks problem and then point out some obstacles to generalizing it to the beads problem. The only part of this section that is used later is equation (3), which is a well-known expression for the expectation of a discrete random variable in terms of tail probabilities. This plays a central role in the solution of the beads problem in subsequent sections.

Let $T(n)$ be a random variable whose value is the draw on which the first matching sock is obtained, when n pairs of socks are initially in the dryer. Let $Q(n,i)$ be the probability that $T(n) > i$. Clearly, $Q(n,0) = 1$, and $Q(n,n+1) = 0$ since there must be a match on or before the $(n+1)^{th}$ draw. If no match has been obtained on or before the i^{th} draw, then on the $(i+1)^{th}$ there are $2n-i$ socks left in the dryer and $2(n-i)$ of them do *not* have a mate that has been drawn. Thus, we have a recursive formula for Q:

$$Q(n,i+1) = \frac{2(n-i)}{2n-i} Q(n,i). \tag{2}$$

Now we can calculate the expected number of socks directly from Q. Recall that if T is a random variable with non-negative integer values, probability mass function $p(i) := P(T=i)$ and tail probabilities $Q(i) := P(T > i)$, then the expected value of T is:

$$E(T) = \sum_{i=1}^{\infty} i\, p(i) = \sum_{i=1}^{\infty} \sum_{j=1}^{i} p(i) = \sum_{j=1}^{\infty} \sum_{i=j}^{\infty} p(i) = \sum_{j=0}^{\infty} Q(j). \tag{3}$$

Accordingly, the expected value of $T(n)$ is

$$E(T(n)) = Q(n,0) + Q(n,1) + \cdots + Q(n,n). \tag{4}$$

$E(T(n))$ is easily evaluated for small values of n. For $n = 1,2,3,4$, the values are respectively $2, 8/3, 16/5, 128/35$. As a matter of fact,

$$E(T(n)) \;=\; \frac{2}{1} \cdot \frac{4}{3} \cdots \frac{2(n-1)}{2n-2} \cdot \frac{2\,n}{2n-1}.$$

This is an immediate consequence of the following proposition, since the right hand side of the equation above can be rewritten as the product that appears on the right hand side of the equation below.

Proposition 2.1.

$$E(T(n)) \;=\; 4^n \binom{2n}{n}^{-1}.$$

Proof. We derive the proposition from the "unexpected" identity 5.20 of [3]:

$$\sum_{k=0}^{n} \binom{n+k}{k} 2^{-k} = 2^n. \tag{5}$$

Multiply both sides of (5) by $2^n \binom{2n}{n}^{-1}$. In the resulting sum on the left, make the substitution $i = n - k$ to get the sum from $i = 0$ to n of the terms:

$$2^i \binom{2n-i}{n-i} \binom{2n}{n}^{-1} = 2^i \binom{n}{i} \binom{2n}{i}^{-1} = Q(n,i). \tag{6}$$

The first equality in line (6) is a consequence of the identity $\binom{r}{n}\binom{n}{i} = \binom{r}{i}\binom{r-i}{n-i}$ (see [3], 5.21); the second is immediate from the recursion in equation (2).

The solution to the socks problem we have just given does not generalize easily to the beads problem. At any stage prior to completing the first k-set there are many possible holdings, each arising with different probability and each leading to a different probability of success on the following draw. For example, after drawing $k-1$ beads, one might be a single draw away from a complete set, or one might have a single bead from each of several different sets. Thus, it is difficult to write a recursive formula for the tail probabilities. It is quite a surprise that a formula as simple as (1) works for all values of k and c.

3. Ascending walks in lattice boxes

We use the following notation. $\mathbb{N}$ is the set of non-negative integers. If $\alpha, \beta \in \mathbb{N}^n$, $\alpha \le \beta$ means $\alpha_i \le \beta_i$, for all i. The *degree* of α, denoted $|\alpha|$, is $\alpha_1 + \alpha_2 + \cdots + \alpha_n$.

Fix $\mathbf{r} \in \mathbb{N}^n$ and let $R := \{\, \alpha \in \mathbb{N}^n \mid \alpha \le \mathbf{r} \,\}$. Let $r := |\mathbf{r}|$; this is the largest degree of any point in R. If $\alpha \in R$, the *opposite point of* α is $\mathbf{r} - \alpha$. The *faces* of R are the subsets $F \subseteq R$ determined by conditions of the form $\alpha_i = 0$ or $\alpha_i = r_i$. The *codimension of* F is the number of conditions required to define it; the *dimension* is n minus the codimension. Thus, R itself is the unique face of codimension 0. The faces other than R are called *proper*. The *opposite face* of F is $\mathbf{r} - F = \{\, \mathbf{r} - \alpha \mid \alpha \in F \,\}$. A face is called *distal* if it contains $\mathbf{r}$ and *proximal* if it contains $\mathbf{0}$.

If $X \subseteq \mathbb{N}^n$, *the degree-d part of* X, denoted $X^{(d)}$, is $\{\, \alpha \in X \mid |\alpha| = d \,\}$. For each $d = 0, 1, \ldots, r$, let $A(d)$ be a random variable with values in $R^{(d)}$, the degree d part of R. Assume that for all d, $A(d) \le A(d+1)$, or equivalently that $A(d+1)$ is the result of increasing one of the coordinates of $A(d)$ by 1. Then, A describes a random walk that starts at $\mathbf{0}$ and at each step moves to a new position by increasing one coordinate by 1, but never leaving R. The walk terminates after r steps, when of necessity it arrives at $\mathbf{r}$. We call this an *ascending walk in* R.

Set $p(\alpha) := P(A(|\alpha|) = \alpha)$. This is the probability that the particle is at position α at time $d = |\alpha|$. The restriction of p to $R^{(d)}$ is the probability mass function of $A(d)$. Note that if $A(d_0)$ is in a distal face, then $A(d)$ belongs to the same face for all $d > d_0$. If $X \subseteq R$, then we call $\sum\{\, p(\alpha) \mid \alpha \in X \,\}$ the *weight of* X and denote it $W(X)$. Our results rest on the computation of the weights of various sets, especially unions of faces and the degree d parts of such unions. Note that $W(R) = 1 + r$.

Let T be a random variable whose value is the smallest d so that $A(d)$ belongs to a proper distal face, i.e., T is the number of steps required to arrive for the first time in a proper distal face. The cumulative distribution function of T is conveniently described in terms of weights. Let $U(1)$ be the union of the proper distal faces. Since the particle is on a distal face at time d if and only if it has arrived at a distal face at a time $\leq d$, the cumulative distribution is:

$$\begin{aligned} G(d) = P(T \leq d) &= \sum \{\, p(\alpha) \mid d = |\alpha| \text{ \& } \alpha \text{ is in a proper distal face} \,\} \\ &= W\Big(U(1)^{(d)}\Big). \end{aligned}$$

Using equation (3),

$$E(T) \;=\; \sum_{d=0}^{r} (1 - G(d)) \;=\; W(R) - W(U(1)). \tag{7}$$

The same approach allows us to express the expected number of draws until other quotas are fulfilled in terms of the weights of subsets of R. For example, let T_c be the smallest d so that $A(d)$ belongs to a distal face of codimension $\geq c$. (In the bead model, this is the number of draws required to complete c sets of beads.) Let $U(c)$ be the union of the distal faces of codimension $\geq c$, and let $U^\circ(c) = U(c) \setminus U(c+1)$. Then the cumulative distribution function of T_c is

$$G_c(d) = P(T_c \leq d) = W\big(U(c)^{(d)}\big).$$

Accordingly,

$$E(T_c) = W(R) - W(U(c)), \tag{8}$$

and

$$E(T_{c+1}) = E(T_c) + W(U^\circ(c)). \tag{9}$$

4. Weights of faces in urn models

The urn model presented in the introduction can be treated in this framework. Indeed, an even more general situation can be considered. Suppose the urn contains n sets of beads and that the i^{th} set contains r_i beads. Then we can attempt to find the expected value of $T =$ the number of draws required to complete one set. This is given (in theory) by formula (7). But to evaluate this, we need to know $p(\alpha)$ and we need to be able to do a summation over a potentially very complex index set. In this section, we address these problems.

The first problem is easy. Since $p(\alpha)$ is the probability of having taken α_i of the $\mathbf{r}_i$ beads of color i, when $|\alpha|$ beads are drawn from a set of r beads, it is the multidimensional hypergeometric probability mass function. To express this, let us use the following notation: for $\mathbf{r}, \alpha \in \mathbb{N}^n$, let

$$\binom{\mathbf{r}}{\alpha} := \prod_{i=1}^{n} \binom{r_i}{\alpha_i}.$$

Then

$$p(\alpha) = \binom{\mathbf{r}}{\alpha}\binom{|\mathbf{r}|}{|\alpha|}^{-1}. \tag{10}$$

The second problem is harder, but for the particular p that we are considering some simplification is possible by exploiting the fact that summing over distal faces is the same as summing over proximal faces, since $p(\alpha) = p(\mathbf{r} - \alpha)$. We will compute the weights of the proximal faces. Fix i and consider the face $F_{\{i\}}$ determined by $\alpha_i = 0$. Let $\mathbf{s}$ be the vector obtained by replacing the i^{th} component of $\mathbf{r}$ by 0. Then if $\alpha \in F_{\{i\}}$,

$$\binom{\mathbf{r}}{\alpha} = \binom{\mathbf{s}}{\alpha}.$$

Let $s = |\mathbf{s}| = r - r_i$ and $d = |\alpha|$. Let q be the probability function for an ascending walk in $F_{\{i\}}$ (considered as a subset of $\mathbb{N}^{n-1}$). Exchanging the denominators in (10), we see that:

$$p(\alpha) = \binom{s}{d}\binom{r}{d}^{-1} q(\alpha).$$

Since $\sum\{\, q(\alpha) \mid \alpha \in F_{\{i\}}^{(d)} \,\} = 1$, the weight of $F_{\{i\}}^{(d)}$ (which is calculated from p) is:

$$W\left(F_{\{i\}}^{(d)}\right) = \binom{s}{d}\binom{r}{d}^{-1}.$$

From this, we get

$$W\left(F_{\{i\}}\right) = \sum_{d=0}^{s} \binom{s}{d}\binom{r}{d}^{-1} = \frac{1+r}{1+r-s} = \frac{1+r}{1+r_i}.$$

The second equality in the line above is proved in Problem 1, page 173, of [3]. A quicker approach is as follows: let $S(r,s)$ stand for the sum indexed by d. Then $S(r,s) = 1 + \frac{s}{r}S(r-1,s-1)$, so the equality follows by induction.

The same reasoning that we applied to $F_{\{i\}}$ yields a formula for the weight of a proximal face of any codimension. Let I be a subset of $\{1, 2, \ldots, n\}$, let F_I be the face determined by demanding that $\alpha_i = 0$ for all $i \in I$ and let $r_I = \sum\{\, r_i \mid i \in I \,\}$. Then,

$$W(F_I) = \frac{1+r}{1+r_I}.$$

In order to compute the expectations $e(n,k,c)$, we need to find weights of unions of distal faces. The weight of the union of all proper distal faces of codimension one can be computed from the information we have already adduced using the inclusion-exclusion principle, since the intersections of pairs of distal faces of codimension one are the codimension 2 distal faces $F_{\{i,j\}}$, the intersections of triples are the codimension 3 distal faces, etc. If we wish to compute the weight of the union of all distal faces of codimension $c > 1$, we need to take into account the fact that the codimension of an intersection of such faces is not determined by the number of faces intersecting. In this case, a more powerful counting principle is needed. It turns out that binomial inversion (a special case of Möbius inversion) suffices. We will elaborate in the next section, as needed.

5. Sets of equal size

In this section, we complete the proof of the main theorem by applying the theory developed above to the special case when $r_i = k$ for all i. Under this assumption, all the distal faces of any given codimension are equivalent. We will write $w(n,k,c)$ for the weight of a face of codimension c. The last result of the previous section tells us:

$$w(n,k,c) = \frac{1+n\,k}{1+c\,k}. \tag{11}$$

In the following, we will make use of the partial fraction expansion of the reciprocal of a polynomial with evenly spaced roots:

$$\frac{1}{x(x+k)(x+2k)\cdots(x+nk)} = \frac{1}{k^n\,n!}\sum_{i=0}^{n}(-1)^i\binom{n}{i}\frac{1}{x+ik}, \tag{12}$$

for which see Remark 8.5, page 68 of [4]. (A special case appears as equation 5.41 in [3].)

Lemma 5.1. *Let $w^\circ(n,k,c)$ be the weight of the set of integer points in a distal face of codimension c that do not lie in any distal face of higher codimension. Then:*

$$w^\circ(n,k,c) \quad = \quad \frac{k^{n-c}\,(n-c)!}{\big(1+c\,k\big)\cdots\big(1+(n-1)\,k\big)} \tag{13}$$

Proof. For simplicity, rather than working with codimension, c, we work with dimension, $\delta = n-c$. Put $v(n,k,\delta) := w(n,k,n-\delta)$ and $v^\circ(n,k,\delta) := w^\circ(n,k,n-\delta)$. We compute $v^\circ(n,k,\delta)$ as below, each step being justified in the comments following the derivation:

$$\begin{aligned}
v^\circ(n,k,\delta) \quad &= \quad \sum_{j=0}^{\delta}(-1)^{\delta+j}\binom{\delta}{j}v(n,k,j) && (14)\\
&= \quad \sum_{j=0}^{\delta}(-1)^{\delta+j}\binom{\delta}{j}\frac{1+n\,k}{1+(n-j)\,k} && (15)\\
&= \quad \sum_{i=0}^{\delta}(-1)^{i}\binom{\delta}{i}\frac{1+n\,k}{1+(n-\delta)\,k+i\,k} && (16)\\
&= \quad \frac{k^\delta\,\delta!}{\big(1+(n-\delta)\,k\big)\cdots\big(1+(n-1)\,k\big)}. && (17)
\end{aligned}$$

Line (14) follows by binomial inversion (see [3], page 192) from:

$$v(n,k,\delta) \quad = \quad \sum_{j=0}^{\delta}\binom{\delta}{j}v^\circ(n,k,j),$$

which itself is simply the result of counting the distal faces of a δ-dimensional hypercube. Line (15) is from (11) and the definition of v. Line (16) is obtained by the substitution $i = \delta - j$ and the last line is from (12).

Let us record some immediate corollaries. First, let $U^\circ(n,k,c)$ be the set of all integer points that lie is some distal face of codimension c but not in any distal face of higher codimension. Since there are $\binom{n}{c}$ such faces, we have:

$$W\big(U^\circ(n,k,c)\big) = \binom{n}{c} w^\circ(n,k,c) = \frac{n!}{c!}\frac{k^{n-c}}{\big(1+c\,k\big)\cdots\big(1+(n-1)\,k\big)}. \tag{18}$$

Second, note that $e(n,k,1) = w^\circ(n,k,0)$. Hence

$$e(n,k,1) = \frac{k^n\, n!}{(1+k)(1+2\,k)\cdots(1+(n-1)\,k)}. \tag{19}$$

This is the $c=1$ case of our main theorem. We now restate and prove the full result, with a minor reformatting of the product on the right of (1):

Theorem 5.1. *Let $n \geq c \geq 1$ and $k \geq 1$ be integers and let $e(n,k,c)$ denote the expected number of draws required to remove c entire k-sets from an urn containing n distinct k-sets. Then*

$$e(n,k,c) \quad = \quad \frac{n!}{(c-1)!}\frac{k^{n-c+1}}{(1+ck)(1+(c+1)k)\cdots(1+(n-1)k)}. \tag{20}$$

Proof. Let $R(n,k,c)$ be the right hand side of equation (20). We show that $e(n,k,c) = R(n,k,c)$ by induction on c. The $c=1$ case is equation (19). Fix $c>1$ and assume that $e(n,k,c) = R(n,k,c)$ for all $k \geq 0$ and $n \geq c$. Note that $e(n,k,c) = k\,B$ and $W\big(U^\circ(n,k,c)\big) = \frac{1}{c}\,B$, where

$$B := \frac{n!}{(c-1)!}\frac{k^{n-c}}{(1+ck)\cdots(1+(n-1)k)}.$$

By equation (9), $e(n,k,c+1) = e(n,k,c) + W\big(U^\circ(n,k,c)\big)$. Thus

$$e(n,k,c+1) = \left(\frac{1+c\,k}{c}\right)B = R(n,k,c+1).$$

6. Asymptotics

We close by looking at some asymptotic properties of the function $e(n,k,1)$. Fix k. For large n, what proportion of the items are removed at the time of the expected first match? That is, how does the ratio $e(n,k,1)/(nk)$ behave as $n \to \infty$? To answer this, we examine the reciprocal of this ratio. For integers $0 \leq m < n$, let

$$R_m^n := \prod_{i=m+1}^{n}\Big(1+\frac{1}{k\,i}\Big).$$

A straightforward calculation shows that, $R_0^{n-1} = n\,k/e(n,k,1)$.

Let $S_m^n := \frac{1}{m+1} + \frac{1}{m+2} + \cdots + \frac{1}{n}$. Recall that

$$\ln(n+1) - \ln(m+1) < S_m^n < \ln(n) - \ln(m).$$

Let $\omega(m) = \ln(1 + \frac{1/k}{m})$. Then, if $0 < x < \frac{1}{k\,m}$,

$$e^{\frac{\omega(m)}{k\,m}x} < 1 + x < e^x.$$

If we let $x = \frac{1}{k(m+1)}, \frac{1}{k(m+2)}, \ldots, \frac{1}{n}$, multiply and take logarithms, we get:

$$\frac{\omega(m)}{m}\frac{S_m^n}{k} < \ln(R_m^n) < \frac{S_m^n}{k}.$$

Hence

$$\frac{\omega(m)}{m}\frac{\ln(n+1) - \ln(m+1)}{k} < \ln(R_m^n) < \frac{\ln(n) - \ln(m)}{k}.$$

Applying the exponential function, we get:

$$\left(1 + \frac{1/k}{m}\right)^{\frac{1}{m}} \left(\frac{n+1}{m+1}\right)^{\frac{1}{k}} < R_m^n < \left(\frac{n}{m}\right)^{\frac{1}{k}}.$$

From this, we can conclude that there are positive constants A and B such that

$$A\,n^{\frac{1}{k}} < R_0^n < B\,n^{\frac{1}{k}}.$$

Accordingly, $e(n,k,1)/(nk)$ is between two positive constant multiples of $\sqrt[k]{1/n}$.

References

[1] Anderson, K., Sobel, M. and Uppuluri, V. (1982). Quota fulfillment times. *The Canadian Journal of Statistics* **10,** 73–88.

[2] gondolin.blogspot.com/2009/04/socks-problem.html

[3] Graham, R., Knuth, D. and Patashnik, O. (1994). *Concrete mathematics: a foundation for computer science*, 2nd edn. Addison-Wesley, Reading, MA.

[4] Larsen, M. (2007). *Summa Summarum.* Canadian Mathematical Society. A K Peters, Ltd., Wellesley, MA.