Outline Version 0.2

Rhett S. Robinson Natalie Hudson Dana Walker
Lee Gilman

July 9 2003

1 Introduction

The coupon collector's problem (CCP) considers the probability of having one copy of all n coupons after t trials, if a trial is defined as collecting one of the n coupons with uniform probability.

The baseball card collector's problem (BCCP) extends this problem to the case where instead of collecting one coupon at a time, r distinct coupons are collected at a time.

This problem may be further generalized by considering it a problem of three variables. The general problem is defined for (n, r, c) where n is the total number of cards to collect, r is the number of cards drawn per trial, and c is the number of copies each card in the set you want. The problem is then to describe the probability at time t of having c copies of all n cards ($P_{(n,r,c)}(t)$).

2 Coupon Collector's Problem

First, a review of a solution to the CCP using inclusion-exclusion.

Define S as the set of all distributions of coupons. Then the size of S, $N(S) = n^t$. Define E_k as the event that k coupons are not collected. Then the size of E_1, $N(E_1) = \binom{n}{1}(n-1)^t$, as there are $\binom{n}{1}$ ways to choose which coupon to not collect, $(n-1)$ coupons from which to choose, and t choices to be made . Similarly, $N(E_2) = \binom{n}{2}(n-2)^t$, and in general $N(E_k) = \binom{n}{k}(n-k)^t$.

We are interested in counting the number of ways that every coupon is collected. We may count this by counting all possible ways of collecting coupons and subtracting off all ways of not collecting at least one coupon. By inclusion-exclusion, we get:

$$\sum_{k=0}^{n}(-1)^k \binom{n}{k}(n-k)^t$$

as the number of ways of collecting t coupons so we get at least one copy of every coupon.

Thus to find the probability of this occurring we divide by the size of our event space, S:

$$P_{(n,1,1)} = \sum_{k=0}^{n} (-1)^k \binom{n}{k} \left(1 - \frac{k}{n}\right)^t$$

3 Baseball Card Collector's Problem

A generalization of the CCP is to consider the case of collecting r distinct coupons at a time. Thus for each trial, we can collect one of $\binom{n}{r}$ packs of coupons at a time, which bears a resemblance to collecting packs of baseball cards. Thus $N(S) = \binom{n}{r}^t$.

If we again define E_k as the event that k coupons are not collected, we find that there are $\binom{n}{k}$ ways of choosing which k coupons to not collect, and then $\binom{n-k}{r}^t$ ways of choosing t packs, none of which contain any of the k coupons.

Again applying the principle of inclusion-exclusion, we determine that the probability of collecting all n coupons after t trials is equal to

$$\frac{\sum_{k=0}^{n} (-1)^k \binom{n}{k} \binom{n-k}{r}^t}{\binom{n}{r}^t}$$

which equals

$$P_{(n,r,1)} = \sum_{k=0}^{n} (-1)^k \binom{n}{k} \left(\frac{\binom{n-k}{r}}{\binom{n}{r}}\right)^t$$

3.1 Asymptotics

We investigate the $\frac{\binom{n-k}{r}}{\binom{n}{r}}$ term in the equation.

$$\frac{\binom{n-k}{r}}{\binom{n}{r}} = \frac{(n-k)...(n-k-r+1)}{n(n-1)...(n-r+1)}$$

$$= (\frac{n-k}{n})^r \frac{(1-\frac{0}{n-k})...(1-\frac{r-1}{n-k})}{(1-\frac{0}{n})...(1-\frac{r-1}{n})}$$

Since $1 - \frac{x}{n} \simeq e^{\frac{-x}{n}}$, we have:

$$= (\frac{n-k}{n})^r \frac{e^{-\frac{k}{n}} e^{-\frac{k+1}{n}} ... e^{-\frac{k+r-1}{n}}}{e^{-\frac{1}{n}} e^{-\frac{2}{n}} ... e^{-\frac{r-1}{n}}}$$

Note that for small k, $(\frac{n-k}{n})^r$ goes toward 1 and this term drops out. We sum the exponents in the numerator from 1 to $r-1$.

$$= -\sum_{i=0}^{r-1} \frac{k+i}{n}$$

$$= -(\frac{k}{n}\sum_{i=0}^{r-1} 1 + \frac{1}{n}\sum_{i=0}^{r-1} i)$$
$$= -\frac{kr}{n} - \frac{(r-1)(r-2)}{2n}$$

We now sum the exponents of the denominator.

$$= -\sum_{i=1}^{r-1} \frac{i}{n}$$
$$= -\frac{(r-1)(r-2)}{2n}$$

Subtract this from the sum of the exponents in the numerator.

$$= -\frac{kr}{n} - \frac{(r-1)(r-2)}{2n} - (-\frac{(r-1)(r-2)}{2n})$$
$$= -\frac{kr}{n}$$

So we have $e^{\frac{-kr}{n}}$. Let's put this result into our original probability formula.

$$\sum_{k=0}^{n} (-1)^k \binom{n}{k} e^{\frac{-krt}{n}}$$

For k sufficiently small, the k's drop out and we have:

$$(1 - e^{\frac{-rt}{n}})^n$$

(if k is large, the sum tends to go toward 0.) Now, let's set $e^{\frac{-rt}{n}} \simeq \frac{x}{n}$. Then we have:

$$(1 - e^{\frac{-rt}{n}})^n \simeq e^{-x}$$

If $x = e^{-c}$, we have:

$$e^{\frac{-rt}{n}} = \frac{e^{-c}}{n}$$

Then, $(1 - e^{\frac{-rt}{n}})^n \simeq e^{-e^{-c}}$

$$e^{\frac{-rt}{n}} = \frac{e^{-c}}{n}$$
$$\log(e^{\frac{-rt}{n}}) = \log(\frac{e^{-c}}{n})$$
$$\log(e^{\frac{-rt}{n}}) = \log(e^{-c}) - \log(n)$$
$$\frac{-rt}{n} = -c - \log(n)$$
$$\frac{-rt}{n} + \log(n) = -c$$

So we get the following formula for c:

$$\frac{rt}{n} - \log(n) = c$$

4 Greedy Baseball Card Collector's Problem ($c = 2$)

Next we consider the further generalization of the baseball card collector's problem to what could be termed a "greedy" baseball card collector's problem, in that we want more than one copy of each card.

Specifically here we consider the case $c = 2$, where we want at least two copies of each card.

4.1 Inclusion-Exclusion

Here we will develop a formulation for the probability we seek with inclusion-exclusion. To do so, we will first count all the ways of having fewer than 2 copies of 0 cards, then subtract all the ways of having fewer than 2 copies of 1 card, and so on up to $n - r$. The value $n - r$ is chosen because after sufficiently many draws, we are guaranteed t o have at least 2 copies of each $n - r$ cards. Thus we only need to consider the ways of not h aving at least c copies of i cards for $0 \leq i \leq n - r$.

For the general case of counting the number of ways of not getting at least 2 copies of i cards, we consider all the possible ways of getting 0 copies of some subset of those i cards and 1 copy for the remaining cards.

For this purpose, define

$$P_{s,u} = \text{The number of ways of picking t packs with 0 copies of s} - \text{u cards and 1 copy of u cards}$$

Then for each $s \in \{0, 1, \ldots, n - r\}$, there are $\binom{n}{s}$ ways of selec ting s cards. For each of these sets of size s, we wish to consider how many ways there are to get less than 2 copies of each card in the set. Thus you can get 0 copies of all s cards, 0 copi es of $s - 1$ cards and 1 copy of 1 card, and in general 0 copies of $s - u$ cards and 1 copy of u c ards for $0 \leq u \leq s$. Note also that for each u, there are $\binom{s}{u}$ ways of picking the u card s we get 1 copy of.

Applying the principle of inclusion-exclusion to our problem, we thus get that t he probability of having all n cards at time t is given by the following:

$$P_{n,r,2}(t) = \sum_{s=0}^{n-r} (-1)^s \binom{n}{s} \sum_{u=0}^{s} \binom{s}{u} P_{s,u}$$

We thus have merely to find a formula for $P_{s,u}$. Unfortunately, this is somewhat complicated as it involves the ways of partition ing u.

Let us consider the case where s and u are fixed. As mentioned, there are $\binom{s}{u}$ ways of choosing which cards among the s can be chosen to have 1 copy of each card. The posi tions of these cards do not change the number of ways t packs can be chosen with 0 copies of $s - u$ car ds and 1 copy of u cards.

Because we may place the u cards in more than one pack, we must consider all p ossible partitions of u and consider placing each part into a separate pack. Thus to calculate $P_{s,u}$ we will iterate over all partitions of u and consider the ways of placing the u cards into packs according to the parts.

We will assume that the partitions of u may be indexed in some fashion, and define the v^{th} partition of u as $Pn(u,v)$, and thus the k^{th} part of the partition as $Pn(u,v)_k$. Also, $|Pn(u,v)|$ will be the number of parts in that partition of u. Finally, define $NO(Pn(u,v),p)$ as the number of occurrences of p in $Pn(u,v)$. Our formula is thus:

$$P_{s,u} = \sum_{v=1}^{p(u)} \binom{t}{|Pn(u,v)|} \frac{(|Pn(u,v)|)!}{\prod_{!p\in Pn(u,v)} NO(Pn(u,v),p)!} \binom{n-s}{r}^{t-|Pn(u,v)|}$$
$$\prod_{k=1}^{|Pn(u,v)|} \binom{\sum_{j=k+1}^{|Pn(u,v)|} Pn(u,v)_j}{Pn(u,v)_k} \binom{n-s}{r-Pn(u,v)_k}$$

4.2 Markov Chains

We may also approach this problem using Markov chains. We define the state (c_1, c_2) to be the state in which we have c_1 cards with at least 1 copy and c_2 cards with at least 2 copies.

4.2.1 Transition Matrix

We can define the probability of moving from state (j_1, j_2) to state (i_1, i_2) as

$$Pr((j_1,j_2) \to (i_1,i_2)) = \frac{\binom{n-j_1}{i_1-j_1}\binom{j_1-j_2}{i_2-j_2}\binom{j_2}{r-(i_1+i_2-j_1-j_2)}}{\binom{n}{r}}$$

An explanation of the formula: $i_1 - j_1$ is the number of cards you will see for the first time in this pack and you are choosing them from $n - j_1$ total cards you have not seen; you are also choosing $i_2 - j_2$ cards that you have seen before but of which you are now getting your first duplicate, and you are picking those from the $j_1 - j_2$ cards you have seen only once; finally, that leaves $r - (i_1 - j_1 + i_2 - j_2)$ cards to fill the pack of r cards, and you are choosing these from the j_2 cards you have already seen at least twice before. To get a probability, we divide by the $\binom{n}{r}$ possible packs.

Using this formula, we can create a transition matrix giving the probabilities of moving from one state to the next.

First we need a way of converting from a pair (i,j) to a pair of states $((i_1,i_2),(j_1,j_2))$ by which to index our matrix. Since certain pairs do not make sense, we do not need to consider a full n^2 matrix: for a given pair (j_1, j_2), we must have that $j_2 \leq j_1$, so we only need to consider $\left(n^2+3n+2\right)/2$ entries.

Working closely with Maple, we developed the following two procedures for converting from a matrix index i to a state (i_1, i_2)

```
majind := proc (n)
   local i, sum;
   sum := 0: i := 0:
   while sum < n do
      sum := sum + i :
      i := i + 1 :
   end do :
   i - 2 ;
end proc ;
```

and

```
minind := proc (n)
   n - sum (i, i=1..majind(n)) - 1;
end proc;
```

These two procedures are used to map a specific row or column to a coordinate: $n \mapsto (\text{majind}(n), \text{minind}(n))$. Thus the transition matrix is assigned entries according to:

$$P_{ij} = Pr((\text{majind}(j), \text{minind}(j)) \rightarrow (\text{majind}(i), \text{minind}(i)))$$

Note: this matrix is 1-indexed to agree with the way in which Maple addresses elements in a matrix.

4.3 Eigenvalues and Eigenvectors

Of particular interest to our problem are the eigenvectors for the nonzero eigenvalues. The eigenvectors contain the binomial coefficients and are dependent on the multiplicity of the eigenvalue, not the actual eigenvalue.
i is the index of the eigenvector starting from the bottom row (we set i of the bottom row equal to zero.
x is the multiplicity of the eigenvalue λ.
For the $\left(\left(\frac{n^2+3n+2}{2}\right) - i\right)$th to the $\left(\frac{n^2+3n+2}{2}\right)$th entries in the eigenvector, the entry is given by

$$(-1)^i \binom{x-1}{i}$$

That is,

i	*eigenvector*
$\left(\frac{n^2+3n+2}{2}\right) - 1$	0
$\left(\frac{n^2+3n+2}{2}\right) - 2$	0
$\vdots$	$\vdots$
$x-1$	$(-1)^i \binom{x-1}{x-1}$
$x-2*$	$(-1)^i \binom{x-1}{x-2}$
$\vdots$	$\vdots$
i	$(-1)^i \binom{x-1}{i}$
$\vdots$	$\vdots$
0	$\binom{x-1}{0}$

Proof: Shannon is working on it.

5 First Order Preimages

Calculations for the first order preimages are given as follows. Note that for this vector, we index the vector from the top instead of the bottom and we begin indexing at 1 instead of zero.

i	$preimage$
1	0
2	0
$\vdots$	$\vdots$
$\frac{(n-1)^2+3(n-1)+2}{2}-(x-2)$	$(-1)^i\left(\lambda r\frac{1}{n-x}\right)^{-1}\binom{x-2}{x-2}$
$\frac{(n-1)^2+3(n-1)+2}{2}-(x-3)$	$(-1)^i\left(\lambda r\frac{1}{n-x}\right)^{-1}\binom{x-2}{x-3}$
$\vdots$	$\vdots$
$\frac{(n-1)^2+3(n-1)+2}{2}-(x-t), 2\leq t\leq x$	$(-1)^i\left(\lambda r\frac{1}{n-x}\right)^{-1}\binom{x-2}{x-t}$
$\vdots$	$\vdots$
$\frac{(n-1)^2+3(n-1)+2}{2}$	$(-1)^i\left(\lambda r\frac{1}{n-x}\right)^{-1}\binom{x-2}{0}$
$\frac{(n-1)^2+3(n-1)+2}{2}+1$	0
$\vdots$	$\vdots$
$\frac{(n-1)^2+3(n-1)+2}{2}+(n-x)+1$	0
$\vdots$	$Free$
$\frac{n^2+3n+2}{2}$	$Variables$

The general form for the $\left(frac(n-1)^2+3(n-1)+22-(x-2)\right)$th to the $\left(\frac{(n-1)^2+3(n-1)+2}{2}\right)$th entries in the vector are given by:

$$(-1)^i\left(\lambda r\frac{1}{n-x}\right)^{-1}\binom{x-2}{x-t}$$

so that $2\leq t\leq x$.

5.1 Second Order Preimages

The second order preimages appear to be of the form:

i	$preimage$
1	0
2	0
$\vdots$	$\vdots$
$\frac{(n-2)^2+3(n-2)+2}{2}-(x-3)$	$(-1)^i\left(\lambda c\sum_{z=1}^{n-r-1}\frac{(z+1)^2}{z(z+2)}\right)^{-1}\binom{x-3}{x-3}$
$\frac{(n-2)^2+3(n-2)+2}{2}-(x-4)$	$(-1)^i\left(\lambda c\sum_{z=1}^{n-r-1}\frac{(z+1)^2}{z(z+2)}\right)^{-1}\binom{x-3}{x-4}$
$\vdots$	$\vdots$
$\frac{(n-2)^2+3(n-2)+2}{2}-(x-t), 3\leq t\leq x$	$(-1)^i\left(\lambda c\sum_{z=1}^{n-r-1}\frac{(z+1)^2}{z(z+2)}\right)^{-1}\binom{x-3}{x-t}$
$\vdots$	$\vdots$
$\frac{(n-2)^2+3(n-2)+2}{2}$	$(-1)^i\left(\lambda c\sum_{z=1}^{n-r-1}\frac{(z+1)^2}{z(z+2)}\right)^{-1}\binom{x-3}{0}$
$\frac{(n-2)^2+3(n-2)+2}{2}+1$	$Zeros$
$\vdots$	$Free$
$\frac{n^2+3n+2}{2}$	$Variables$

The general form for the $\left(frac(n-2)^2+3(n-2)+22-(x-3)\right)$th to the $\left(\frac{(n-2)^2+3(n-2)+2}{2}\right)$th entries in the vector are given by:

$$(-1)^i\left(\lambda c\sum_{z=1}^{n-r-1}\frac{(z+1)^2}{z(z+2)}\right)^{-1}\binom{x-3}{x-t}$$

so that $3\leq t\leq x$.
c is unknown. It may either be a constant, a variable, or some value dependent on n and/or r.

5.2 Jordan Block Form

Discuss how eigenvalues and eigenvectors fill in, etc.

5.3 Double-sum formula

We are interested in the probability of going from having 0 cards to having 2 copies of all n cards, so we are interested in the $((n^2+3n+2)/2, 1)$ entry of PJ^tP^{-1}. Having found the Jordan form of the matrix, we can find this probability by

$$\begin{aligned}&\sum_{i=0}^{n-r}\sum_{k=0}^{mult(\lambda_i)-1} c_{\lambda_i,k}\cdot\lambda_i^{t-k}\cdot\binom{t}{k}\\ &\text{where } \lambda_i = \frac{\binom{i}{r}}{\binom{n}{r}}\end{aligned}\tag{1}$$

which represents what the answer will look like for sufficiently large t (here, the t must be larger than the largest block of 0 eigenvalues so that those blocks play no role in the probability.

After having computed many $c_{\lambda,k}$ values, we compared the actual probabilities with those found by using

$$\sum_{\lambda_i} c_{\lambda_i,\mathrm{mult}(\lambda_i)-1}\cdot\lambda_i^{t-\mathrm{mult}(\lambda_i)+1}\cdot\binom{t}{\mathrm{mult}(\lambda_i)-1}$$

and

$$\sum_{\lambda_i}\sum_{k=\mathrm{mult}(lambda)-2}^{\mathrm{mult}(\lambda_i)-1} c_{\lambda_i,k}\cdot\lambda_i^{t-k}\cdot\binom{t}{k}$$

Our results indicate that the first is a good upper bound, while the second is an extremely close fit (see Figure 1).

Figure 1: Comparing actual probabilities with those using a subset of $c_{lambda,k}$ values

We have done some analysis of the resulting values of $c_{\lambda,k}$, and have spotted some patterns.

By looking at Equation 1 and at **ref to $P_{(n,r,2)}(t)$ here**, we were able to determine that for the case when $k = \mathrm{mult}(\lambda_i) - 1$

$$c_{\lambda_i,k} = \frac{(-1)^i\binom{n}{i}\binom{n-i}{r-1}^i i!}{\binom{n}{r}^i}$$

and for the case when $k = \mathrm{mult}(\lambda_i) - 2$

$$c_{\lambda_i,k} = \frac{(-1)^i\binom{n}{i}\left((i-1)!\binom{i}{2}\binom{n-i}{r-2}\binom{n-i}{r-1}^{i-2}+\binom{i}{1}(i-1)!\binom{n-i}{r-1}^{i-1}\right)}{\binom{n}{r}^{i-1}}$$

This should lead to some nice asymptotics when we get around to it (hopefully).

6 Misc

6.1 r=2, r=3

We've written some specific versions for $r = 2$ and $r = 3$. We can fill those in here if there's any particular reason to do so.

Okay, here's what we've got for $r = 2$:

$$P_{(n,2,2)} = \frac{\sum_{s=0}^{n-2} (-1)^s \binom{n}{s} \sum_{u=0}^{s} \binom{s}{u} \sum_{n_2=0}^{\lfloor \frac{u}{2} \rfloor} \frac{\binom{t}{u-n_2}(u-n_2)!\binom{n-s}{r}^{t-u+n_2} u!(n-s)^{u-2n_2}}{n_2!(u-2n_2)!2^{n_2}}}{\binom{n}{r}^t}$$

The $r = 3$ case is a bit more complex, but follows a similar pattern. I'll refrain from giving it for just now.

7 Future Work

What's left to be done. This paper, for example.

8 Conclusion

It's been a lot of fun.