Department of Mathematical Sciences,
Clemson University.

http://people.clemson.edu/ lgilman

Calculations and Asymptotics of the Baseball Card Collector Problem

Lee G. Gilman

lgilman@clemson.edu

Objectives

- Statement of problem
- Probability Formula for the single copy (non greedy) case
- Sample results for $n = 50$, $r = 3$
- Derivation of the asymptotic formula

The Problem

We are given a set S of n distinct elements and during each time interval we take a subset of S of size r

We want to know the probability that we will have seen all n elements at some time $t > 0$.

Actually, we are interested in when t is sufficiently large so that the probability is sufficiently close to 1. The probability will never be exactly 1 (i.e., the same subset of r elements may appear for every time interval.)

Inclusion/Exclusion Derivation of Probability Formula

For each trial, we collect one of $\binom{n}{r}$ subsets. For t trials, we have $\binom{n}{r}^t$ total possible subsets that we can choose from.

Using inclusion-exclusion, we wish to count all of the ways we can choose subsets (given by $\binom{n}{r}^t$ for t trials) and then subtract all such possibilities where less than n cards are collected.

Let's say that k elements do not appear. For each trial, there are $\binom{n}{k}$ ways of selecting which k elements are not to appear. For t trials, we have $\binom{n-k}{r}^t$ ways of choosing t subsets, none of which contain any of the k elements. Using inclusion-exclusion, we get the following result:

The probability Formula:

For the non-greedy baseball card collector problem, the probability of having seen all n cards at time t is

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

Here, n is the size of the entire set S, r is the size of the subsets of S that are taken at each time interval. t is the number of time intervals, and k is a dummy variable that ranges from 0 to n.

Computations for $n = 50$

The probability formula was entered into Maple and run for $n = 50$, $n = 100$, $n = 150$, $n = 200$, and $n = 250$.

On the next two slides are the results for $n = 50$.

Computations for $n = 50$

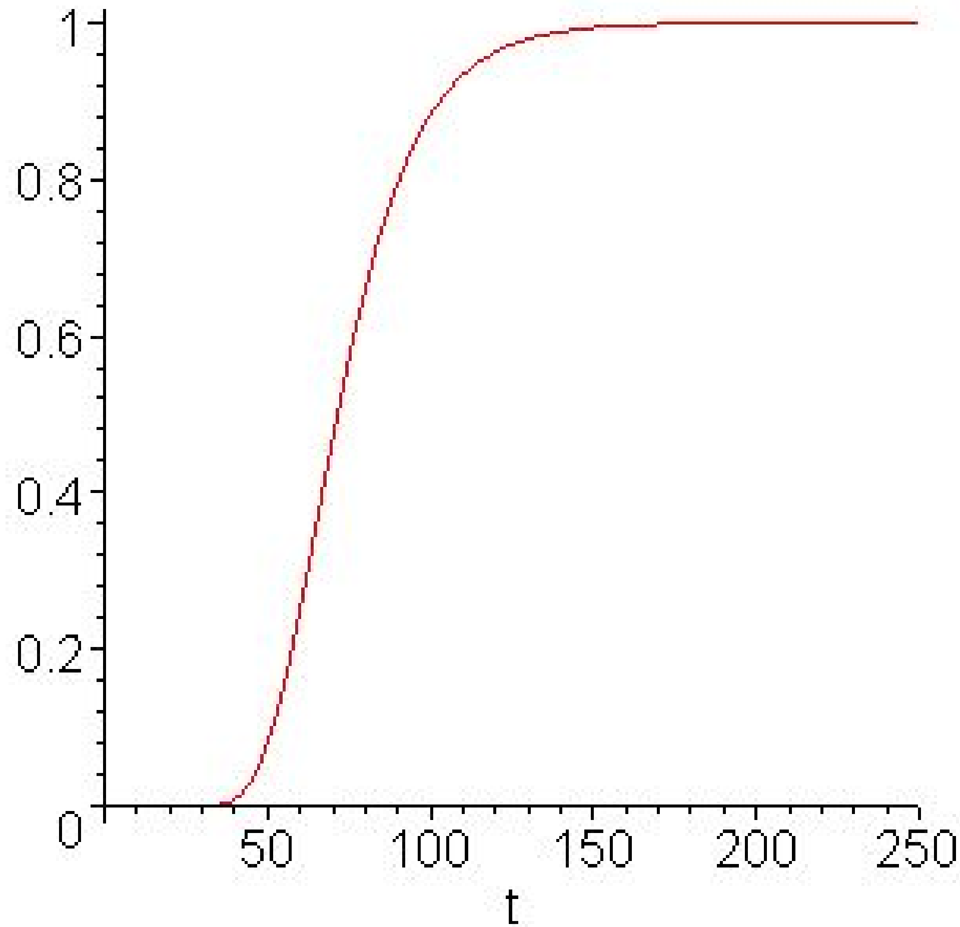
t	$Pr(t)$
0	0
10	0
20	0
30	0
40	0.006
50	0.081
60	0.273
70	0.506
80	0.696

Computations for $n = 50$

t	$Pr(t)$
90	0.824
100	0.902
110	0.946
120	0.971
130	0.984
140	0.991
150	0.995
160	0.997
170	0.999
180	0.999

Graph for $n = 50$

Maple was used to graph the probability curve for $n = 50$. This graph is shown below.



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:

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

Asymptotics

We now have this formula:

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

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 $r-1$.

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

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

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

Asymptotics

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 $e^{\frac{-kr}{n}}$.

Asymptotics

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:

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

Asymptotics

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

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

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

Asymptotics

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

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

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

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

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

So we get the following formula for c:

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

References

[1] R.B. Bapat and T.E.S. Raghavan *Nonnegative Matrices and Applications*, Cambridge University Press, 1997.

I would also like to thank Shannon Purvis for helping me format this presentation and getting it prepared.