ATTACKING VARIATIONS OF THE COUPON COLLECTOR PROBLEM WITH MAPLE

Dr. Joel C. Fowler
Mathematics Department
Kennesaw State University, Marietta Campus
1100 South Marietta Parkway
Marietta, GA 30060-2896
jfowle60@kennesaw.edu

The Coupon Collector Problem concerns the number of draws, from n equally likely coupon types, that are needed to collect a full set of coupons. We consider extensions and variations of that classic problem. Because the complexity of the computations requires technology, Maple is used.

In the classic Coupon Collector's Problem, a collection of n equally likely coupon types are repeatedly drawn from in order to accumulate a full set of n. The original problem is to determine the expected number of draws needed to reach a full set of n coupons. In modern times, this problem could be likened to repeated visits to a fast-food restaurant that is giving away one toy at random, from a set of n different toys, with each meal purchased. In this setting the question would be to find the mean number of meals needed to obtain a full set of the n toys. The elegant answer, which is well known to be approximately n ln(n), relies on properties of geometric random variables and sums of random variables.

Here we consider more detailed questions associated with the Coupon Collector's Problem beyond the expected value. The mathematics employed, and Maple usage, is at the level of a senior undergraduate mathematics major project. We consider exact probabilities in several ways. Our direct formulas involve inclusion-exclusion and Stirling numbers, while another method is recursive in nature.

We let n be the number of coupons, k the number of draws at random from the set of coupons, and d the size of a desired set of different coupons within the set of n.

Let P(d, k, n) = the probability of having exactly d different coupons after k draws.

Then, working recursively, we have that P(d, k, n) can be found by considering two cases for the n-th draw. One is that d different coupons were already obtained within the first (k - 1) draws and no new coupon was obtained on the k-th. The other is that (d - 1) different coupons were obtained within the first (k - 1) draws, and k-th draw is a new coupon. This leads to:

$$(1) \qquad P(d,\ k,\ n) = P(d,\ k-1,\ n)\frac{d}{n} + P(d-1,\ k-1,\ n)\frac{n-d+1}{n} \quad ,$$

with the initial conditions:

$$(2) \quad P(1,\ k,\ n) = \frac{n \cdot 1^n}{n^k} = \frac{1}{n^{k-1}},\ \text{for}\ k \geq 1 \quad \text{, and P(d, k, n) = 0, for d >k or d > n.}$$

The Maple code to implement this recursion is straightforward:

```
> ProbRecur := (d,k,n) -> if d>min(k,n) then 0 else if d=1 then 1/(n^(k-1)) else
d*ProbRecur(d,k-1,n)/n + (n-d+1)*ProbRecur(d-1,k-1,n)/n  end if end if;
```

This generates values for smaller values of the parameters very efficiently, but other methods are needed to generate larger values if we wish to study the distribution. A direct inclusion-exclusion approach, based on n sets, each of which excludes one of the n coupon types, yields:

$$(3) \qquad P(d, k, n) = \frac{\binom{n}{d}}{n^k} \sum_{i=0}^{d} (-1)^i \binom{d}{i} (d-i)^k \quad .$$

Again a Maple implementation is easy:

```
> ProbDirect := (d,k,n)->evalf(Sum((-1)^i*binomial(d,i)*(d-i)^k,i=0..d )*binomial(n,d)/n^k
);
```

P(d, k, n) is essentially a cumulative probability in k, since it doesn't presume the set of d was completed on the k-th turn. For our study of the distribution it will be helpful to have the probability function as well.

Let

PE(d, k, n) = Probability that a set of d, from n, is completed on the k-th turn

= P(d, k, n) - P(d, k-1, n).

By using (3) and manipulating the difference we find:

$$(4) \qquad PE(d, k, n) = \frac{\binom{n}{d}}{n^k} \sum_{i=1}^{d-1} (-1)^{i+1} \binom{d}{i} i\,(d-i)^{k-1}$$

Our last approach, before studying the distribution, is via Stirling Numbers of the Second Kind. It's equivalent to both the recursive and direct computation methods. But observing the connection does open the door on all of the known properties of the Stirling Numbers.

We use the notation S(i, j) for the Stirling Numbers of the Second Kind. Recall that:

S(i, j) = the number of partitions of i different objects into j unordered, non-empty locations.

The number of ways that k turns can result in exactly d coupons represented is the same as the number of ways that the k turns can be distributed onto the d coupon types, with each coupon type non-empty. The only difference is that the Stirling Numbers presume unordered locations, while the d coupon types are distinct, and therefore ordered. Hence we have:

$$(5) \qquad P(d, k, n) = \frac{\binom{n}{d}}{n^k} S(k, d)\, k!$$

From this perspective we note that (3) essentially embodies one of the known computational formulas for the Stirling Numbers of the Second Kind.

Implementing in Maple, using the built in Stirling Number routines, can be done via:

```
> with(combinat);
> ProbStir := (d,k,n) -> evalf( binomial(n,d)*stirling2(k,d)*d!/n^k );
```

A Stirling form for PE(d, k, n) could also be created, but we'll have little need for it. However one special case, that of a full set of n coupons in k turns, will be useful. So the Stirling form for that case is worth pursuing.

From (1), with d = n, we have

$$P(n, k, n) = P(n, k-1, n) + P(n-1, k-1, n)\frac{1}{n} \quad .$$

Hence

$$PE(n, k, n) = P(n, k, n) - P(n, k-1, n) = P(n-1, k-1, n)\frac{1}{n} \quad .$$

Applying (5) and simplifying yields:

$$(6) \qquad PE(n, k, n) = \frac{(n-1)!}{n^{k-1}} S(k-1, n-1) \ ,$$

which can be implemented in Maple via:

```
> ProbFullExactStir := (k,n) -> evalf( stirling2(k-1, n-1)*(n-1)!/n^(k-1) );
```

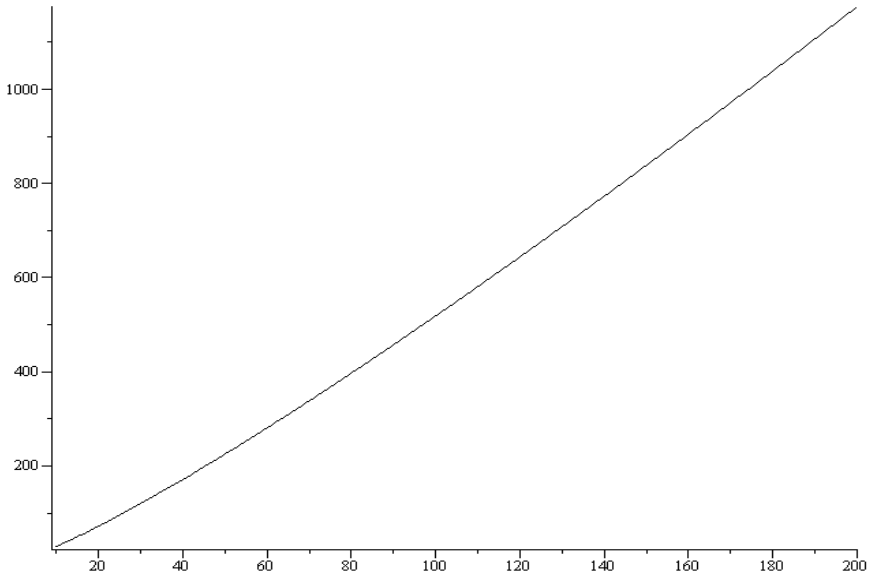
Running time-trials in Maple reveals that the formulas based on the built in Stirling Number routines are, by quite a margin, the fastest. Hence those are used for the computations that follow. And we now consider probabilities only for collecting full collections of n coupons.

We first verify the classic expected value solution to the Coupon Collector Problem numerically. We define a computational approximate mean for via:

```
> CompMean := n -> evalf( sum( j*ProbFullExactStir(j,n),j=n..15*n ) );
```

The limit for the sum of 15*n was arrived at by examining the distribution for values of n up to 200 and informally checking where the probabilities became very close to zero. We then create vectors we can graph and work with via::

```
> XVec := < seq(n,n=10..200) >; Yvec := <vseq(CompMean(n), n=10..200) >;
> pointplot( XVec,YVec, connect=true);
```



A very slight curvature is evident. After trying various curve fitting models involving powers and logs we find the best is:

```
> Fit(a*n*ln(n)+b*n, XVec, YVec, n, output=[leastsquaresfunction]);
```

(7) [0.9940 n ln(n) + 0.6098 n]

The dominant term is quite close to n ln(n), which is the classic answer. It is arrived at analytically as follows. Let

$$X = \text{ total number of draws for a complete set of n}$$

$$= X_1 + X_2 + X_3 + \ldots + X_n,$$

where Xi is the waiting time from the acquisition of the (i - 1)-st new coupon to the i-th new coupon. Each Xi is a geometric random variable with parameter $p = \frac{n - i + 1}{n}$.

Hence each has mean $\frac{1}{p} = \frac{n}{n - i + 1}$. Summing from 1 to n then yields:

$$E(X) = \frac{n}{n} + \frac{n}{n-1} + \frac{n}{n-2} + \ldots + \frac{n}{1} = n \cdot \sum_{i=1}^{n} \frac{1}{i} \approx n \ln(n) .$$

Using ln(n) for a partial sum of the Harmonic Series is a bit sloppy. In fact, a standard result from series is that

$$\lim_{n \to \infty} \left(\sum_{i=1}^{n} \frac{1}{i} - \ln(n) \right) = \gamma, \text{ Euler's constant.}$$

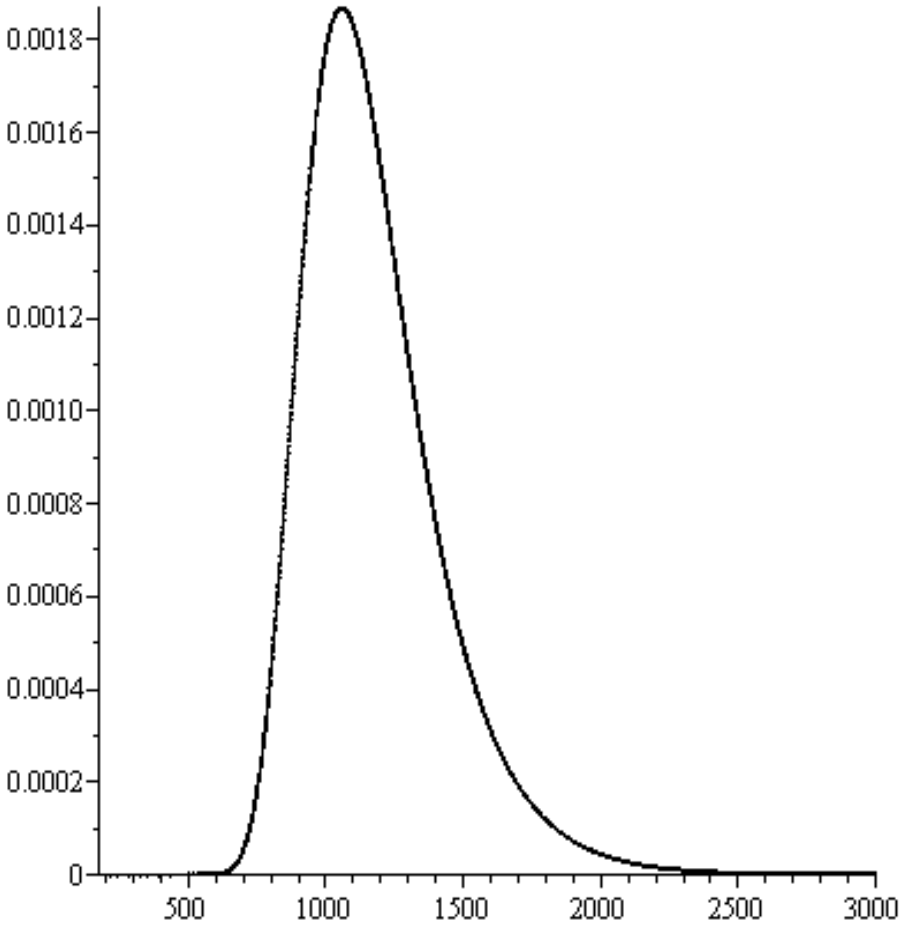
Hence a more refined estimate for the expected value would be

$$n \cdot \sum_{i=1}^{n} \frac{1}{i} \approx n (\ln(n) + \gamma) = n \ln(n) + \gamma n .$$

Comparing this with our regression equation, (7), explains the second, nontrivial, term beyond the n ln(n). The .61 coefficient of n is, in fact, an experimental approximation of Euler's constant, which is ~ .5772 .

We can easily examine the probability distribution using Maple's graphing capabilities.

```
> pointplot({seq([k, ProbFullExactStir(k, 200)], k = 200 .. 3000)}, symbol = circle, symbolsize = 5)
```



We see that the distribution is unimodal and skewed. Motivated by the above graph, our first task will be to prove that the distribution is unimodal.

To address increases and decreases in the sequence of probabilities, we consider the ratio of successive terms. By using (6) and simplifying we have:

(8) $$\frac{PE(n,\ k+1,\ n)}{PE(n,\ k,\ n)} = \frac{1}{n} \cdot \frac{S(k,\ n-1)}{S(k-1,\ n-1)}$$

Since $S(n-1, n-1) = 1$ and $S(n, n-1) = \binom{n}{2}$, we have that this ratio is $\frac{n-1}{2}$ for $k = n$, which is the smallest value of k for non-zero probability. So the ratio is > 1, for $n > 3$.

Hence PE(n, k, n) is initially increasing in k. And we already know that P(n, k, n) tends to 0 for large k, since it is a probability function. Hence this ratio is initially greater than 1 (i.e. PE(n, k, n) is increasing there), and, at some large value of k, is less than 1 (i.e. PE(n, k, n) is decreasing there). Unimodality would be implied by a uniform decrease in this ratio as k increases, from the values greater than 1 to those less than 1.

Since $\frac{1}{n}$ is fixed it suffices to show that $\frac{S(k, n-1)}{S(k-1, n-1)}$ is decreasing in k.

Note that $\frac{S(k, n-1)}{S(k-1, n-1)} > \frac{S(k+1, n-1)}{S(k, n-1)}$ if and only if $S(k, n-1)^2 > S(k+1, n-1)\,S(k-1, n-1)$. However, this is precisely the condition that a sequence be log-concave, and it is well known that the Stirling Numbers are log-concave (in both parameters). Hence we have established that the distribution is unimodal.

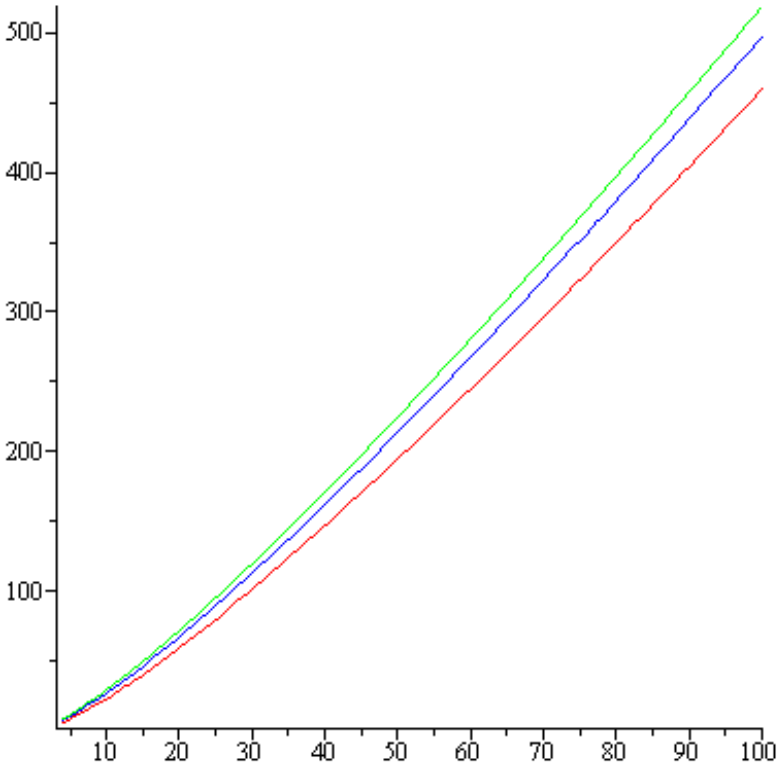
It is interesting to study the mean, mode , and median as n increases. We write two short Maple procedures to find these.

```
> CCPercentile := proc(x,n)
      Z := RandomVariable(Normal( n*ln(n),n*sqrt(ln(n)-1))):
      Value := floor( Percentile(Z,x) ):
          while ProbStir(n,Value,n) > x/100 do Value := Value-1 end do:
          while ProbStir(n,Value,n) < x/100 do Value := Value+1 end do:
      Value;
end proc;
```

```
> CCMode := proc(n)
          Value := floor( 1.01*CCMean(n) ):
    while ProbFullExactStir(Value,n)> ProbFullExactStir(Value+1,n) do Value := Value-1
    end do:
  Value+1;
end proc;
```

We can now graph the mean, median, and mode over a range of values of n.

```
> plot( [ [ seq([i,CCMode(i)],i=4..100)], [ seq([i,CCPercentile(50,i)],i=4..100)], [
seq([i,CCMean(i)],i=4..100)] ], color=[red,blue,green]);
```



The order of the three statistics in the graph is mode < median < mean. It does appear that the three statistics are pulling apart as n increases, which is consistent with a skew distribution.

The mode is particularly interesting. We proceed computationally as with the mean by performing a regression on the data for n versus the mode. For n ranging from 10 through 200 the best regression fit, via Maple, is:

$$1.0039*n*\ln(n)-0.0267*n$$

This is quite suggestive that the mode tends toward the simple expression $n \ln(n)$. We examine this hypothesis analytically.

Since the distribution is unimodal, the mode will occur near the value of k for which the expression in (8) is close to 1. So we seek the value of k for which:

(9) $$S(k, n-1) \approx n \cdot S(k-1, n-1) \ .$$

The standard recurrence for the Stirling Numbers of the Second Kind yields:

$$S(k, n-1) = (n-1)\cdot S(k-1, n-1) + S(k-1, n-2) \ .$$

Using this recurrence together with (9) produces:

$$S(k-1, n-1) \approx S(k-1, n-2) \ .$$

Since the Stirling numbers are unimodal, we seek the relationship between n and k for which they achieve a maximum across a row. This is also a well-known result. Asymptotically, the Stirling Numbers of the Second Kind achieve their maximum value in a row at the row number divided by its natural log. For us this gives, as an approximation:

$$\frac{k}{\ln(k)} \approx n \quad \text{or} \quad k \approx n\ln(k) \ .$$

While this doesn't solve explicitly for k in terms of n, we may iterate the expression to find an approximation for k in term of n:

$$\begin{aligned} k &\approx n\ln(k) \\ &\approx n\ln(n\ln(k)) = n\ln(n) + n\ln(\ln k) \\ &\approx n\ln(n) + n\ln(\ln(n\ln k)) = n\ln(n) + n\ln(\ln(n) + \ln(\ln k)) \ . \end{aligned}$$

This could be carried further, but we already have what was desired, plus additional information. A rough approximation for the mode for large n is, in fact, $n\ln(n)$, and a somewhat more refined approximation would be $n\ln(n) + n\ln(\ln(n))$.

A final, remaining, problem that we have not solved is the behavior of the median. Although we have no proof, the experimental evidence from the distribution places it between the mean and mode, perhaps at $n\ln(n) + cn$, for some constant c between 0 and Euler's Constant..