Filling a Penny Album

Shiyong Lu
Steven Skiena
Department of Computer Science
State University of New York
Stony Brook, NY 11794-4400
{shiyong—skiena}@cs.sunysb.edu

October 19, 1999

Most coin collectors can trace their interest in the hobby back to that childhood day when they were given an empty coin album and encouraged to fill it up. This typically blue, typically cardboard album contained slots the size of the lowest denomination coin in circulation (such as the United States one cent piece or 'penny'), with each slot labeled with the year in which that coin was minted. Filling the album requires collecting one coin of each mintage year.

The first few days of owning an album are always marked by great success, with many slots quickly getting filled. As the days pass, progress gets slower and slower, as we keep getting repeated instances of previously collected years. The excitement builds as the number of slots gets whittled down one by one. Filling in the last slot marks the satisfying end of a quest.

Can we predict how many coins we will be expected to see before we complete an album? Determining an answer requires modeling the process by which coins are removed from circulation and then solving an interesting variant of the famous coupon collector's problem. In this article, we try to determine the number of United States pennies one must assay to construct a complete collection of Lincoln Memorial Cents.

How Often Do I Get a 1959 Penny?

The probability of seeing a coin minted in a given year in circulation is a function of mintage, age, and collector pressure. The *mintage* is the number of coins issued in a given year. Mintages vary widely as a function of the demand for new coins in circulation. Certain years have high mintage, certain years low mintage.

The *age* of a coin is the number of years since it was minted. Older coins become scarcer as they get lost from circulation. The older a coin is, the more people that have handled it, and hence the more likely it is to disappear behind a sofa or down a drain.

Year	Sampled	Predicted	Year	Sampled	Predicted	Year	Sampled	Predicted
1939	1	0	1966	11	8.8	1982	73	97.3
1940	1	0	1967	15	12.5	1983	76	84.7
1941	1	0	1968	16	20.4	1984	69	83.6
1950	1	0	1969	19	24.5	1985	70	68.2
1952	1	0	1970	17	24.2	1986	74	57.1
1955	3	0	1971	22	24.2	1987	78	62.5
1956	1	0	1972	29	27.6	1988	71	75.9
1957	2	0	1973	15	35.9	1989	89	86.3
1958	1	0	1974	41	42.9	1990	65	82.5
1959	6	6.5	1975	47	49.3	1991	68	66.8
1960	11	10.0	1976	35	45.1	1992	63	66.7
1961	3	9.0	1977	40	44.9	1993	114	90.9
1962	8	8.8	1978	40	52.2	1994	129	104.7
1963	10	9.5	1979	44	55.2	1995	140	106.5
1964	25	24.8	1980	60	69.8	1996	172	105.6
1965	7	5.9	1981	67	73.2	1997	49	75.8

Figure 1: Distribution of Dates in $20 Worth of Pennies, as Sampled and Predicted.

The final factor is *collector pressure*. Sufficiently old coins become uncommon as collectors perceive interest and value. Also, coins with obsolete designs tend to be removed from circulation by collectors. For example, in 1959, the United States introduced an image of the Lincoln Memorial on the back of each penny, replacing the 'Wheatback" penny reverse which had been in use since 1909. Today, it is a fairly unusual occurrence to find a wheatback penny, as collectors have tossed them aside for years.

We propose a simple exponential decay model to predict the frequency of circulating coins which have not been subject to collector pressure. Let M_i denote the number of coins minted i years ago. We assume that any coin has a probability of p to be retained in circulation (i.e. not lost) during a given year. Therefore, the number of coins minted i years ago which are still in circulation should be $p^i \cdot M_i$.

We let p denote the *decay coefficient* of the coin denomination. But what is the value of p for United States pennies? To help answer this question, on April 20, 1998 we withdrew 2,000 pennies (i.e. $20 worth) from the Stony Brook branch of the Teachers Federal Credit Union. A summary of what we found appears in the first column of Figure 1. Only 12 of the 2000 pennies were wheatbacks (pre-1959), and so for simplicity we will disregard the existence of these coins in the remainder of the article. The mintage figures of U.S. Lincoln Cents for each year from 1959-1997 appear in standard numismatic references such as R. S. Yeoman's *A Guidebook of United States Coins*, Western Publishing, Racine Wisconsin.

Estimating the Decay Coefficient

Given a value for the decay coefficient, we can compute the proportion of the coins in circulation from each mintage year and the number of representatives from each mintage

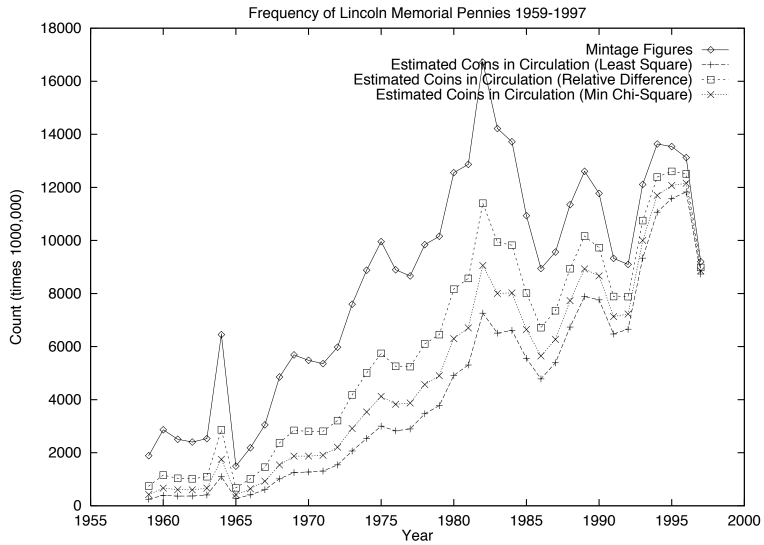

Figure 2: Distribution of United States Pennies in Circulation

year that would be expected in a sample of 2000 coins (actually 1988 coins in our sample). To estimate the decay coefficient, we compute the value of p that minimizes the difference between our actual sample and the expected count. Our prefered criterion is the sum of the squares of the relative differences (the difference divided by the expected count). We did not use the figures from 1997 in minimizing the sum of squares, since we did not have full mintage figure for that year. The resulting estimate of p is 0.976. The corresponding estimate of coins in circulation is more than 234 million. Figure 2 gives the mintage figure for each year along with the estimated coins in circulation from the exponential decay model. The second column of Figure 1 gives the expected counts corresponding to the exponential decay model; there is a good agreement with the sample.

It is natural to wonder how sensitive our analyses are to the optimization criterion used to estimate p, and to assess the sampling variability in our estimate. To access sampling variability, we generate 100 random sets of pennies by sampling with replacement from our actual sample. We estimate p for each sample and use these estimates to assess sampling variability. This is a statistical technique known as the bootstrap. We also consider two alternative criteria: ordinary least squares (minimizing the sum of squared differences between observed and expected counts) and minimum chi-squared (minimizing the sum of the squares of the differences divided by the square root of the expected count). The results are summarized in Figure 3. Sampling variability is quite small but the results are sensitive to criteria. Figure 2 presents the coins in circulation estimate for all of the criteria. Small differences in p leads to substantial differences in the number of coins estimated to be circulating. We use the relative difference criterion because it seems to fit best for the rare, early dates which are crucial to fill the penny album.

Criteria	Mean	Min	Max	Std. Dev	Predicted Pennies in Circulation
Least Square	0.9492	0.9407	0.0959	0.004	165,521,321,515
Relative Diff	0.9764	0.9673	0.9856	0.003	234,894,161,052
Min Chi-Square	0.9624	0.9568	0.9715	0.003	194,969,089,381

Figure 3: Variation in optimal decay coefficient p over 100 random coin samples, according to three different optimization criteria.

The Weighted Coupon Collectors Problem

Suppose a coupon collector needs to collect a complete set of n coupons (each with a distinct number from 1 to n) in order to get a prize. Assuming that one coupon is enclosed within a box of (say) cereal, what is the expected number X of cereal boxes the collector must buy in order to capture the prize? In the best case, buying n boxes will suffice, however we will likely have to buy substantially more because of duplicates.

This so-called *coupon collector's problem* has a simple and elegant solution when the coupons occur with equal probability, i.e. the probability of the next box containing coupon i is $p_i = 1/n$. To solve this problem, we partition the process of collecting coupons into n phases – each phase we wind up with a new distinct coupon that is not in the collection yet. Let X_i denote the number of boxes bought in the i^{th} phase. For the first phase, all coupons are new, so $X_1 = 1$. For the second phase, the next box contains a duplicate with probability $1/n$ and a new distinct coupon with probability $(n-1)/n$. We keep selecting boxes until we find a distinct coupon. The number of boxes is described by the geometric distribution with probability of success $(n-1)/n$, so the expected value of X_2 is $n/(n-1)$. Once phase two completes we continue to phase three. In general, the probability that the next box contains a new distinct coupon (which indicates the end of phase i) is $(n-i+1)/n$, and hence the expected number of boxes in this phase is $n/(n-i+1)$. The total waiting time is the sum of the phases. Then the expected waiting time is

$$\sum_{i=1}^{n} \frac{n}{n-i+1} = n \sum_{i=1}^{n} \frac{1}{i} \approx n \ln n$$

where ln is the natural logrithm and the last approximation is valid for large n. Our collector should budget to buy approximately $n \ln n$ cereal boxes.

However, our penny collecting problem is complicated by the fact that the number of pennies in circulation from different mintage years is not the same. In the general coupon collectors problem, the probabilities $p_1, p_2, ..., p_n$ of the n coupons are not equal. The solution to this more general case was first published by Herman Von Schelling in 1954 in the *American Math. Monthly.* For the weighted case, he shows that the expected number of boxes required to get a full set is

$$\sum_{1 \le i_1 \le n} \frac{1}{p_{i_1}} - \sum_{1 \le i_1 < i_2 \le n} \frac{1}{p_{i_1} + p_{i_2}} + \cdots + (-1)^{n-1} \frac{1}{p_1 + p_2 + \cdots + p_n}$$

Though we don't repeat the proof here, it is interesting to look a bit more at this result. Suppose that we divide the general coupon collecting procedure for this general problem into i phases such that phase i is only concerned with collecting the i^{th} coupon. Since the i^{th} coupon occurs with probability p_i, the expected time for phase i is $1/p_i$. If each phase was independent, the total coupon collection time would be $\sum_{i=1}^{i=n} 1/p_i$. This is the first term of our formula. It is clear that this is too big, because at any point we accept any unseen coupons. How should we correct for this overestimate? Von Shelling's formula is analogous to the inclusion-exclusion formula for the size of a set intersection. The second term corrects the first term by substracting off a term that corresponds to the expected time when each pair is sampled. The result is an underestimate so the third term is used to correct the underestimate. The terms alternate to produce the final result.

Computational Issues

Evaluating the general coupon collectors formula requires generating all 2^n subsets of coupon-types, since there is one term for each subset in Von Shelling's formula. These combinations can be exhaustively generated. However, since the number of such combinations grows exponentially, the formula can be practically computed only for small values of n.

Since this formula requires time exponential in n, we propose a less expensive way to compute a lower bound on the expected number of pennies. Let us define a $m \leq n$ such that we can solve instances of the weighted of size m in a reasonable amount of time. With current technology, a reasonable value might be $m \approx 20$.

Our approach to a lower bound is to partition the n items into m groups, and compute the exact weighted coupon collector time to get at least one representative from each group. The results of any such partitioning will give us a lower bound on the coupon collector solution of the n items. Our approach to the partitioning was to sort the coupons in increasing order of probability, and take as our partition $m-1$ singleton groups comprising the $m-1$ smallest probabilities, and one group consisting of the $n-m+1$ items of highest probability. The intuition behind this partitioning is that it is much more likely that the last coin inserted into the album is a low probability representative than a high probability one.

Our approach to the upper bound is again based on partitioning into groups. We divide the n coupons into $k = \lceil \frac{n}{m-1} \rceil$ groups, and divide the collecting procedure into k phases: In phase i, we seek to collect one of each of the coupons in group i, and one of the coupons not in group i (We bunch all the other coupon types into another separate group, say complementary group i' and consider it as one type of coupon with the possibility of the sum of all the possibilities of coupon types in group i'). After all of the phases are completed, we have a complete set of coupons. However, in phase i, those coupons in the complementary group i' will also be covered by one of the other phases. This fact indicates the result would be an upper bound.

	Least Square	Relative Diff	Min Chi-Square
Expected Number	1313.763	683.888	945.994
Lower Bound	1313.762	683.867	945.990
Upper Bound	1437.081	809.299	1068.397

Figure 4: Projected number of pennies needed under three difference optimization criteria.

A Small Study

Now we can put all these results together to determine bounds on the number of Lincoln Memorial pennies needed to fill the album.

The years 1959-1997 comprise 39 dates of pennies, whose estimated frequency in circulation is given in Figure 2. Figure 4 gives the value of our upper and lower bounds for all three optimization criteria, using a group size of $m = 20$. In quest of tighter bounds, we did exhaustive calculations as well, each of which ran for about 2.3 CPU days on the fastest machine available in our department. In fact, under our preferred measure we expect to see 683.888 coins before filling our penny album, which is extremely close to the more quickly computed lower bound.

On January 27, 1999, we went through a set of accumulated pennies to see how long it took to collect all years. In fact, on the 630th coin we obtained the 1962 penny needed to complete our collection – quite in accord with our results.

Our story thus far has focused only on the expected number of coins. In our actual study with real coins, it took 630 coins to get a full set, a bit less but quite close to the expected value. It is natural to wonder how much variability there is in the coin collecting process. Again, the typical coupon collector's problem admits a straightforward discussion because the standard deviation of the number of boxes to get a complete set has an accessible (but complicated) formula. There is no formula for the weighted case. To give some idea, we simulated filling an album using 30 random permutations of the 2000 pennies of our original sample. The average collecting period was 714.1 coins, ranging from a low of 296 to a high of 1413.

Filling a penny album remains an affordable goal for children young and old.

References

[1] W. Feller. *Probability Theory and Its Applications*, volume I. John Wiley & Sons, Inc., New York, 1950.

[2] H.J. Goodwin. On cartophily and motor cars. *Math. Gazette*, 33:169–171, 1949.

[3] M. G. Kendall and B. Babington Smith. Randomness and random sampling numbers. *Journal of the Royal Statistical Society*, 101:147–166, 1938.

[4] F. G. Maunsell. A problem in cartophily. *Math. Gazette*, 29:77, 1939.

[5] Hermann von Schelling. Coupon collecting for unequal probabilities. *American Mathematics Monthly*, 61:306–311, 1954.

Biographies

Shiyong Lu is a Ph.D. student in Computer Science at SUNY Stony Brook. He graduated from the University of Science and Techonology of China at Hefei, P. R. China in July 1993 and came to the United States in August 1996 after earning his M.E. from Academia Sinica at Beijing. Currently he works with his advisors Arthur J. Bernstein and Philip M. Lewis on concurrency in high performance transaction processing and workflow systems. His other interests include model checking e-commerce protocols, algorithm analysis, databases, networkings and web search. He has published several papers in international conferences. His hobbies includes swimming, basketball, tennis and dining. Check out his web page at `http://www.cs.sunysb.edu/~shiyong/`.

Steven Skiena is an Associate Professor of Computer Science at SUNY Stony Brook, and has been interested in coin collecting since being given a penny album as a child. He received his Ph.D. in Computer Science from the University of Illinois at Urbana-Champaign in May 1988. His research interests include combinatorial algorithms and discrete mathematics, and particularly their applications to biology. He is the author of two books: *The Algorithm Design Manual*, published by Springer-Verlag and *Implementing Discrete Mathematics*, published by Perseus Books. Check out his web page at `http://www.cs.sunysb.edu/~skiena/`.