A Generalized Coupon Collector Problem

Seenu S. Reddi
ReddiSS at aol dot com
July 15, 2006

The coupon collector problem can be stated as follows. A person collects n coupons one at a time from a store and the problem is to estimate the average number of tries and related statistics for collecting all n coupons. When all the coupons are equally probable, the average number is asymptotically n ln n + O(n) (see Amy Myers [1]). It has applications to randomized algorithms and one may look up in the internet for articles dealing with this problem.

When the coupons are not equally probable, the problem appears to be formidable and to my knowledge, there are no bounds established on the average number and its statistics. Assume the n coupons 1, 2, .., n have probabilities p_1, p_2, p_3, …, p_n and the problem is to estimate the number of tries to collect all coupons. However the problem becomes amenable to numerical methods once it is formulated as a Markov chain. Define 2^n states S_0, S_1, S_2, .., S_n, $S_{1,2}$, $S_{1,3}$, …, $S_{1,2,3,..,n}$ corresponding to sets {}, {1}, {2}, .., {n}, {1, 2}, {1, 3}, …, {1, 2, 3, .., n} and there will be transitions corresponding to each try. Note the transitions can happen only between states whose number of elements correspond to a number and its Peano's successor. Thus there will be transitions between S_0 to S_1, S_1 to $S_{1,2}$, $S_{1,2}$ to $S_{1,2,3}$, .. but not between S_1 to $S_{1,2,3}$. Define a missing index function MIF(S, T) on states S and T whose elements differ by 1 as the one that gives the integer that corresponds to the missing element in S. Thus MIF($S_{1,2}$, $S_{1,2,3}$) = 3. Define P(S, T) as $p_{MIF(S, T)}$ and P(S, S) as the sum of probabilities of the elements in the set corresponding to S. Set up a matrix M that corresponds to these transitions and start with a row vector V = (1, 0, 0, …, 0). Vectors V*M, V*M*M, …. (where * denotes matrix multiplication) contain the occupational probabilities of states after one transition, two transitions, … From the occupational probability of the last state corresponding to the set {1, 2, .., n}, we can compute the average number of tries to collect all the coupons.

An example will illustrate the process. Assume n = 3 and we choose the states S_0, S_1, S_2, .., S_7 so that its binary representation of the subscript reflects the presence or absence of the elements. Thus we have:

0: (0, 0, 0)	S_0 = {}
1: (0, 0, 1)	S_1 = {1}
2: (0, 1, 0)	S_2 = {2}
3: (0, 1, 1)	S_3 = {1, 2}
4: (1, 0, 0)	S_4 = {3}
5: (1, 0, 1)	S_5 = {1, 3}
6: (1, 1, 0)	S_6 = {2, 3}
7: (1, 1, 1)	S_7 = {1, 2, 3}

and the transition matrix M can be computed for probabilities 0.16, 0.68, 0.16 (corresponding to an approximate Gaussian distribution) as:

$$M = \begin{bmatrix} 0.00 & 0.16 & 0.68 & 0.00 & 0.16 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.16 & 0.00 & 0.68 & 0.00 & 0.16 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.68 & 0.16 & 0.00 & 0.00 & 0.16 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.84 & 0.00 & 0.00 & 0.00 & 0.16 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.16 & 0.16 & 0.68 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.32 & 0.00 & 0.68 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.84 & 0.16 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 \end{bmatrix}$$

V * M gives (0.00 0.16 0.68 0.00 0.16 0.00 0.00 0.00) indicating that the probabilities of occupation for states 1, 2 and 4 as 0.16, 0.68 and 0.16. The following table gives the occupational cumulative probabilities for the final state and we can compute the average length as 9.46.

3: Avg. Len. = 0.313344	V = 0.104448
4: Avg. Len. = 0.731136	V = 0.208896
5: Avg. Len. = 1.21577	V = 0.305824
6: Avg. Len. = 1.75222	V = 0.395231
7: Avg. Len. = 2.32258	V = 0.476711
8: Avg. Len. = 2.90773	V = 0.549855
9: Avg. Len. = 3.49062	V = 0.614621
10: Avg. Len. = 4.05767	V = 0.671326
11: Avg. Len. = 4.599	V = 0.720537
12: Avg. Len. = 5.10799	V = 0.762953
13: Avg. Len. = 5.58072	V = 0.799317
14: Avg. Len. = 6.01536	V = 0.830363
58: Avg. Len. = 9.45942	V = 0.99991
59: Avg. Len. = 9.46019	V = 0.999932
60: Avg. Len. = 9.46084	V = 0.999943
61: Avg. Len. = 9.4614	V = 0.999952
62: Avg. Len. = 9.46188	V = 0.99996
63: Avg. Len. = 9.46229	V = 0.999966
64: Avg. Len. = 9.46263	V = 0.999971
65: Avg. Len. = 9.46293	V = 0.999976

Note for equal probabilities 1/3, 1/3, 1/3 we get an average length of 5.5. A computer program written for PC can be found at www.rspq.org/pubs/GenCouCol.exe.

[1] Amy Myers, http://www.sju.edu/~amyers/research.html - look for Coupon Collector Problem.