CSE 5311-001 Lab Assignment 1

Due March 3, 2021

Goals:

1. Understanding of coupon collecting.
2. Understanding of enumeration.
3. Understanding of random simulation to verify a probability result.

Requirements:

1. The following paper does a variety of interesting probabilistic analyses:

   P. Flajolet et.al., "Birthday paradox, coupon collectors, caching algorithms and self-organizing search", *Discrete Applied Mathematics 39* (1992), 207-229. (`http://algo.inria.fr/flajolet/Publications/FlGaTh92.pdf`)

   It includes the following formula, which provides the expected number of coupons needed under a ***general*** discrete probability distribution P for m coupons:

   $$\sum_{q=0}^{m-1} (-1)^{m-1-q} \sum_{|J|=q} \frac{1}{1-P_J} \quad \text{(14b)} \qquad \text{where } P_J = \sum_{i \in J} p_i$$

   For m=3 and $(p_1, p_2, p_3) = (a, b, c)$, the paper simplifies (14b) to:

   $$1 - \frac{1}{1-a} - \frac{1}{1-b} - \frac{1}{1-c} + \frac{1}{1-a-b} + \frac{1}{1-b-c} + \frac{1}{1-c-a}.$$

   Your task is write a C program to 1) evaluate (14b) for the ***special*** case that k of the probabilities are the value p and the other m - k probabilities are the value q and 2) implement a simple random simulation of generating coupons for this situation. Your program must compile and execute on `omega.uta.edu`.

   (Formula (1) in `https://web.cs.wpi.edu/~hofri/CCP.pdf` says essentially the same thing as the formula above.)

2. Submit your C code on Canvas before 3:45 p.m. on March 3. Be sure to include comments regarding how to compile and execute your code.

Getting Started:

1. m will not exceed 50. m is at least 2.

2. The input is very simple:

   a. The first input line is m and k. k must be larger than 0 and smaller than m.
   b. The second line will have the value of the probability p for each of the first k coupons. p must be larger than 0 and such that $k \cdot p$ is smaller than 1. Each of the other m - k coupons will have a probability of $q = (1 - k \cdot p)/(m - k)$
   c. The third input line will be the number of times to execute a random simulation.
   d. The fourth input line is a seed for the random number generator (e.g. `srandom()`). (The only significance of the seed is in the reproducibility of the experiments.)

3. Your powerset approach should use $\Theta(m^2)$ time instead of the $\Theta(2^m)$ time needed by the original version of expression (14b). Elementary combinatorics (`https://en.wikipedia.org/wiki/Pascal%27s_triangle`) will be useful. You should output how many times innermost loop(s) execute for the theoretical part. The time for computing binomial coefficients should not be included.