Tutorial 12 solutions.

1. This is a real problem that came from a research project in computational biology a few weeks ago. The CS PhD student hadn't taken 6.042 and so he needs our help. Can you solve the problem?

   It is known that atoms occur in different isotopes. For example, Carbon atoms usually have 6 protons and 6 neutrons, but occasionally they have 7 neutrons. In fact 1.11% of all Carbon atoms have 7 neutrons and 98.89% have 6. From what we know, the appearance of the different isotopes seems to be a random event, unrelated to the functionality of the atom. We can tell which Carbon atoms have an extra neutron only by their atomic weight. The atomic weight of the generic Carbon atom is 12 and of its isotope is 13.

   The research team is studying proteins composed of Carbon, Oxygen, Nitrogen and Hydrogen (the four building blocks of life). Their atomic weights and isotopes are listed below. The second column is the atomic weight and the third column is the frequency of the isotope.

$Hydrogen$	1	$P_{H,1}$
	2	$P_{H,2}$
$Carbon$	12	$P_{C,12}$
	13	$P_{C,13}$
$Nitrogen$	14	$P_{N,14}$
	15	$P_{N,15}$
$Oxygen$	15	$P_{O,15}$
	16	$P_{O,16}$
	17	$P_{O,17}$

   The research team wants to write software that can predict the distribution of atomic weights for any protein. Describe a simple algorithm for computing the weight distribution of any protein that contains n_C carbon atoms, n_O oxygen atoms, n_N nitrogen atoms and n_H hydrogen atoms. [Hint: Use generating functions!]

   Let R_H be the random variable whose value is the atomic weight of a random Hydrogen atom. Similarly, let R_C, R_N and R_O be the random variables for the atomic weight of Carbon, Nitrogen and Oxygen atoms.

We can write down the generating functions for their distributions:

$$\begin{aligned} G_H(z) &= P_{H,1}z + P_{H,2}z^2 \\ G_C(z) &= P_{C,12}z^{12} + P_{C,13}z^{13} \\ G_N(z) &= P_{N,14}z^{14} + P_{N,15}z^{15} \\ G_O(z) &= P_{0,15}z^{15} + P_{O,16}z^{16} + P_{O,17}z^{17} \end{aligned}$$

Then, let R be the random variable representing the atomic weight of a protein with n_C carbon atoms, n_O oxygen atoms, n_N nitrogen atoms and n_H hydrogen atoms. We can write:

$$R = \sum_{i=1}^{n_H} R_{H,i} + \sum_{i=1}^{n_C} R_{C,i} + \sum_{i=1}^{n_N} R_{N,i} + \sum_{i=}^{n_O} R_{O,i}$$

Here the distribution of each hydrogen atom in the protein, $R_{H,i}$, is identical to the distribution of R_H. Similarly, the $R_{C,i}$'s have the same distribution as R_C, the $R_{N,i}$ have the same distribution as R_N and the $R_{O,i}$'s have the same distribution as R_O.

By the product rule, the PGF for R is

$$G(z) = (P_{H,1}z+P_{H,2}z^2)^{n_H}\,(P_{C,12}z^{12}+P_{C,13}z^{13})^{n_C}\,(P_{N,14}z^{14}+P_{N,15}z^{15})^{n_N}\,(P_{0,15}z^{15}+P_{O,16}z^{16}+P_{O,17}z^{17})^{n_O}$$

Hence the probability that a protein molecule with n_H, n_C, n_N, n_O atoms of H, C, N, O respectively has atomic weight k is the coefficient of z^k in the above polynomial.

In fact, the student has now written a program to multiply the polynomials and so the problem is solved!

2. Coupon Collector problem: We are trying to collect baseball cards. Assume there are n different cards. We can buy them one at a time. When we buy a card, it is equally likely that we get any one of the n cards, and purchases are mutually independent. We want to study how many cards we need to buy before we get one each of the n different kinds.

   (a) Suppose we already have k different cards. What is the expected number of cards we need to buy to get a new card?

   Define X_k to be the random variable denoting the number of purchases to get the $(k+1)^{th}$ new card, after getting k different cards.

   After getting k different cards, a new purchase has probability $\frac{n-k}{n}$ of getting a new card. We can view each new purchase as Bernoulli trial with probability of success $p = \frac{n-k}{n}$. Hence the expected number of trials to success is $\frac{1}{p} = \frac{n}{n-k}$. That is,

$$E(X_k) = \frac{n}{n-k}$$

(b) Using the above, what is the expected number of cards we need to buy to get one of each kind?

Since $X_0 = 1$, the number of purchases we need is $1 + \sum_{i=1}^{n-1} X_i$. By linearity of expectation,

$$\begin{aligned} E(1 + \sum_{i=1}^{n-1} X_i) &= 1 + \sum_{i=1}^{n-1} E(X_i) \\ &= 1 + \sum_{i=1}^{n-1} \frac{n}{n-i} \\ &= n \sum_{i=1}^{n} \frac{1}{i} \\ &= n\mathcal{H}_n \sim n \ln n + \gamma n + \Theta(1) \end{aligned}$$

Where γ is Euler's constant.

(c) Suppose we buy $n \ln n + cn$ cards, where $c > 0$ is a real number. What is the probability that we don't have the 1^{st} card?

Let A_1 be the event that we don't have the first card in $n \ln n + cn$ trials.

In any single trial the probability that we don't get A_1 is $(1 - \frac{1}{n})$. Since the trials are independent,

$$\begin{aligned} Pr(A_1) &= (1 - \frac{1}{n})^{n \ln n + cn} \\ &\leq e^{-\frac{1}{n}(n \ln n + cn)} \\ &= e^{-\ln n + c} = \frac{e^{-c}}{n} \end{aligned}$$

(d) What is an upper bound on the probability that we don't have all n cards?

Let A_i be the event that we don't have the i^{th} card after $n \ln n + cn$ trials.

Then the probability that we don't have all the cards is

$$Pr(A_1 \cup \ldots \cup A_n) \leq \sum_{i=1}^{n} Pr(A_i)$$

by Boole's inequality. But this is

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