EECS 70 Discrete Mathematics and Probability Theory
Fall 2014 Anant Sahai

Homework 11

This homework is due November 17, 2014, at 12:00 noon.

1. Section Rollcall!

In your self-grading for this question, give yourself a 10, and write down what you wrote for parts (a) and (b) below as a comment. Put the answers in your written homework as well.

(a) What discussion did you attend on Monday last week? If you did not attend section on that day, please tell us why.

(b) What discussion did you attend on Wednesday last week? If you did not attend section on that day, please tell us why.

2. Biased Coins, Birthday Paradox, and Stirling's Approximation Lab

Up until this point, everything that you have done in the last three virtual labs is something that you could've naturally discovered yourself as something worth trying. The data is speaking directly to the experimentalist in you. However, discovering an actual formula for the shape of this "cliff-face" is something that actually requires a theoretical investigation that is related to counting, Fourier Transforms, and Power Series. Guessing its exact shape is not something that comes very naturally on experimentalist intuition alone.

In this week's lab, we will simply provide you with the right curve and continue from last week's lab on biased coins. Unless specified otherwise, you can assume the same configurations from last week's lab. In other words, the coin is biased with $P(\text{head}) = 0.7$, the number of tosses are $(k = 10, 100, 1000, 4000)$, respectively, and the number of trials is $m = 1000$. Make sure you review the lab solution from Homework 10 before moving on.

In addition, we will also look at the Birthday Paradox and Stirling's Approximation. Please come back to the last three parts of the lab when you are working on Question 8.

For each part, students who want to can choose to completely rewrite the question. Basically, you can come up with your own formulation of how to do a series of experiments that result in the same discoveries. Then, write up the results nicely using plots as appropriate to show what you observed. You can also rewrite the entire lab to take a different path through as long as they convey the key insights aimed at in each part.

Please download the IPython starter code from Piazza or the course webpage, and answer the following questions.

(a) Plot $\int_{-\infty}^{d} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$ overlaid with the normalized cliff-face shapes you had plotted in last week's lab. This integral is related to something called the Error Function. What do you observe?

This is the heart of the Central Limit Theorem as applied to coin tosses.

Hint: Implement the function `normal`, which takes a real number x and returns $\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$. Then, implement the function `integrate_normal(d)`, which integrates the above function from $-\infty$ to d. In Python, you can use `scipy.integrate.quad`.

(b) Now, since you had realized earlier that the cliff-faces and the histograms have some natural relationship with each other, how would you naturally overlay a smooth plot of $\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$ to the normalized histograms. What does this mean?

Hint: There's a parameter for `plt.hist()` you learned in HW7 that you can use to normalize the histogram. Also, use a bin size of 0.2.

(c) Another interesting pattern that you had seen in the previous Virtual Labs was the exponential drop in the frequencies of certain rare events. For an exponential drop, the most interesting thing is to understand the rate of the exponential — or the relevant slope on the Log-Linear plot.

For a coin with probability p of being heads, we are interested in the frequency by which tossing k such coins results in more than ak heads (where a is a number larger than p). We are interested in $p = 0.3, 0.7$ and $a = p + 0.05$, $p + 0.1$. Take $m = 1000$ and plot the natural log of the frequencies these deviations against k (ranging from 10 to 200). Approximately extract the slopes for all four of these.

Compare them in a table against the predictions of the following formula (which we will derive later in the course).

$$D(a||p) = a\ln\frac{a}{p} + (1-a)\ln\frac{1-a}{1-p}.$$

This expression is called the Kullback-Leibler divergence and is also called the relative entropy.

Finally, add $e^{-D(a||p)k}$ to the plots (there should be 4 of these) you have made as straight lines for immediate visual comparison. This straight line is called a "Chernoff Bound" on the probability in question.

What do you observe?

Hint: First, implement the function `KL`, which computes $D(a||p)$ using the given formula. To fit a line in Python, you can use `np.polyfit`.

There will probably be some 0 values, which will mess up this fitting, so you can replace the zeros with 10^{-3}.

(d) During your first week of Charm School (CS), you want to find fellow CS students who have the same birthday. Let's switch gears to an interesting problem studied in Lecture Note 12: the Birthday Paradox. This interesting phenomenon concerns the probability of two people in a group of m people having the same birthdays. This probability is given by

$$P(A^c) = 1 - \frac{365 \times 364 \times \ldots \times (365-m+1)}{365^m} = 1 - \frac{365!}{(365-m)!365^m}$$

where

$$P(A) = \frac{365!}{(365-m)!365^m} = \left(1-\frac{1}{365}\right) \times \left(1-\frac{2}{365}\right) \times \ldots \times \left(1-\frac{m-1}{365}\right)$$

and $P(A)$ is the probability that no two people have the same birthday.

For $m = 10, 23, 50, 60$, randomly generate birthdays by uniformly picking m numbers between 1 and 365. Do this 1000 times for each value of m. Record how many trials have at least 2 same birthdays. Plot this fraction vs. m using a bar chart. What do you observe?

Hint: First, implement the function `has_duplicate`, which returns True if a list contains any repeated element and False otherwise. Then, implement the function `gen_birthday(m)`, which generates random birthday for m people.

(e) We will now calculate the probability of having two people with the same birthday empirically, and plot the result against the expected probability, which is derived in Note 12.

Implement the function `birthday_formula(m)`, which calculates the probability of at least two people having the same birthday among m people.

Plot the empirical result (you can assume 1000 trials) v.s. the analytical result (`birthday_formula(m)`), for $m = [1, 100]$ people. What do you observe about the two curves? What happens at $m = 23$ and $m = 60$? Is this consistent with what we previously knew about the Birthday Paradox?

(f) Now approximate $P(A)$ using Stirling's approximation for $n!$ and plot the approximated $P(A^c) = 1 - P(A)$ as a function of m. Stirling's approximation is given by

$$n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n.$$

Plot the analytical result from the previous part and the approximated result in the same figure. What do you observe?

Hint: Implement the function `birthday_stirling`, which computes the probability that no two people have the same birthday given that there are m birthdays using Stirling's approximation.

That said, don't use Stirling's approximation directly! Simplify your expression after using the approximation as much as possible before you implement the `birthday_stirling` function, or the large values will blow up your computer.

(g) Lastly, let's come back to the problem of counting the number of ways to throw m balls into n bins. Suppose the number of balls in each bin is a nonnegative integer, implement the function `permutation(m,n)`, which generates all possible permutations of throwing m balls into n bins. For example, `permutation(2,3)` should return

`[[0,0,2],[0,1,1],[0,2,0],[1,0,1],[1,1,0],[2,0,0]]`.

How would you change your implementation if we now require each bin to contain a positive number of balls?

Hint: Use recursion. You should have three base cases.

(h) Question 8, part (a)

(i) Question 8, part (h)

(j) Question 8, part (j)

Reminder: When you finish, don't forget to convert the notebook to pdf and merge it with your written homework. Please also zip the `ipynb` file and submit it as `hw11.zip`.

3. Picking CS Classes

The EECS (Elegant Etiquette Charm School) department has d different classes being offered in Fall 2014. These include classes such as dressing etiquette, dining etiquette, and social etiquette, etc. Let's assume that all the classes are equally popular and each class has essentially unlimited seating! Suppose that c students are enrolled this semester and the registration system, EleBEARS (Elegant Bears), requires each student to choose a class s/he plans to attend.

(a) What is the probability that a given student chooses the first class, dressing etiquette?

(b) What is the probability that a given class is chosen by no student?

(c) If there are $d = 20$ classes, what should c be in order for the probability to be at least one half that (at least) two students enroll in the same class?

4. Sock etiquette

In your second week of Charm School you learn that you should only wear matching pair of socks. In each pair, both socks must be of the same color and pattern. But all of them are in one big basket and now you have to take a pair out. Let's say you own n pairs of socks which are all perfectly distinguishable (no two pairs have the same color and pattern). You are now randomly picking one sock after the other without looking at which one you pick.

(a) How many distinct subsets of k socks are there?

(b) How many distinct subsets of k socks which do not contain a pair are there?

(c) What is the probability of forming at least one pair when picking k socks out of the basket?

(d) Now, in a different experiment, suppose there is exactly one sock of each pair in the basket (so there are n socks in the basket) and we sample (with replacement) k socks from the basket. What is the probability that we pick the same sock at least twice in the course of the experiment?

5. Drunk man

Imagine that you have a drunk man moving along the horizontal axis (that stretches from $x = -\infty$ to $x = +\infty$). At time $t = 0$, his position on this axis is $x = 0$. At each time point $t = 1$, $t = 2$, etc., the man moves forward (that is, $x(t+1) = x(t)+1$) with probability 0.5, backward (that is, $x(t+1) = x(t)-1$) with probability 0.3, and stays exactly where he is (that is, $x(t+1) = x(t)$) with probability 0.2.

(a) What are all his possible positions at time t, $t \geq 0$?

(b) Calculate the probability of each possible position at $t = 1$.

(c) Calculate the probability of each possible position at $t = 2$.

(d) Calculate the probability of each possible position at $t = 3$.

(e) If you know the probability of each position at time t, how will you find the probabilities at time $t+1$?

The Drunk Man has regained some control over his movement, and no longer stays in the same spot; he only moves forwards or backwards. More formally, let the Drunk Man's initial position be $x(0) = 0$. Every second, he either moves forward one pace or backwards one pace, *i.e.*, his position at time $t+1$ will be one of $x(t+1) = x(t)+1$ or $x(t+1) = x(t)-1$.

We want to compute the number of paths in which the Drunk Man returns to 0 at time t and it is his first return, *i.e.*, $x(t) = 0$ and $x(s) \neq 0$ for all s where $0 < s < t$. Note, we **no longer** care about probabilities. We are just counting paths here.

(f) How many paths can the Drunk Man take if he returns to 0 at $t = 6$ and it is his first return?

(g) How many paths can the Drunk Man take if he returns to 0 at $t = 7$ and it is his first return?

(h) How many paths can the Drunk Man take if he returns to 0 at $t = 8$ and it is his first return?

(i) How many paths can the Drunk Man take if he returns to 0 at $t = 2n+1$ for $n \in \mathbb{N}$ and it is his first return?

(j) How many paths can the Drunk Man take if he returns to 0 at $t = 2n+2$ for $n \in \mathbb{N}$ and it is his first return? (Hint: read http://en.wikipedia.org/wiki/Catalan_number and use any result there if you need.)

6. An Identity on Integer Partitions

Let n be a positive integer. A partition of n is a way of writing n as a sum of positive integers. Partitions are considered equivalent under permutation of the summands, so that order of the summands does not matter. For example, 3 has exactly 3 partitions:

$$\begin{aligned} 3 &= 3 \\ &= 2+1 \\ &= 1+1+1 \end{aligned}$$

We will represent each partition p as a set of pairs (x,r) where the first element x represents a summand and the second element r is the number of times the summand appears, so that we have $n = \sum_{(x,r)\in p} rx$ for any partition p of n. We denote by $\mathscr{P}(n)$ the set of partitions of integer n. For example:

$$\mathscr{P}(3) = \{\{(3,1)\},\{(2,1),(1,1)\},\{(1,3)\}\}$$

In this problem, we will construct a combinatorial proof of the following identity:

$$\sum_{p\in\mathscr{P}(n)} \prod_{(x,r)\in p} \frac{1}{r!x^r} = 1$$

For example, for $n = 3$, the identity is saying:

$$\left(\frac{1}{1!3^1}\right) + \left(\frac{1}{1!2^1} \times \frac{1}{1!1^1}\right) + \left(\frac{1}{3!1^3}\right) = 1$$

(a) Make sure the above identity works for any $n \leq 5$.

Let σ_n be the set of permutations over $\{1,2,\ldots,n\}$. Let $f \in \sigma_n$. We say that $(x_1\ x_2\ \ldots\ x_k)$ is a cycle of length k of f if and only if $f(x_1) = x_2, f(x_2) = x_3, \ldots, f(x_k) = x_1$. Note that $(x_1\ x_2\ \ldots\ x_k), (x_2\ x_3\ \ldots\ x_k\ x_1), (x_3\ x_4\ \ldots\ x_1\ x_2)$, $\ldots$ all represent the same cycle.

A familiar way to represent a permutation $f \in \sigma_n$ is to explicitly list the mapping $(x, f(x))$ for all $1 \leq x \leq n$. A different way to represent the same permutation is to list all its cycles. Consider Table 1 for an example of this.

x	1	2	3	4	5	6	7
$f(x)$	5	7	4	6	1	3	2

Table 1: A permutation $f \in \sigma_7$. The same permutation can also be represented by 1 cycle of length 3 and 2 cycles of length 2: $(4\ 6\ 3)$, $(2\ 7)$ and $(1\ 5)$.

(b) Suppose we are working in σ_n. How many distinct cycles of length l can one construct?

(c) Let $(l_1,\ldots,l_m)$, $(x_1,\ldots,x_k)$, $(r_1,\ldots,r_k)$ be 3 sequences of positive integers such that:

i. $\sum_{i=1}^{m} l_i = n$,

ii. $\sum_{j=1}^{k} x_j r_j = n$,

iii. For all $j \leq k$, there are exactly r_j distinct $i \leq m$ such that $l_i = x_j$.

How many distinct permutations in σ_n can be represented by a set of m cycles of length $l_1,\ldots,l_m$? Express this number only in terms of n and the rs and xs. Be careful not to over count permutations.

(d) You already know that $|\sigma_n| = n!$ by a simple counting argument. Now, use the previous question to count the elements of σ_n by using their cycle representation in order to prove the above identity.

7. Fibonacci Fashion

You have n accessories in your wardrobe, and you'd like to plan which ones to wear each day for the next t days. As a Charm School student, you know it isn't fashionable to wear the same accessories multiple days in a row. (Note that the same goes for clothing items in general). Therefore, you'd like to plan which accessories to wear each day represented by subsets $S_1, S_2, \ldots, S_t$, where $S_1 \subseteq \{1,2,\ldots,n\}$ and for $2 \leq i \leq t$, $S_i \subseteq \{1,2,\ldots,n\}$ and S_i is disjoint from S_{i-1}.

(a) For $t \geq 1$, prove there are F_{t+2} binary strings of length t with no consecutive zeros (assume the Fibonacci sequence starts with $F_0 = 0$ and $F_1 = 1$).

(b) Use a combinatorial proof to prove the following identity, which, for $t \geq 1$ and $n \geq 0$, gives the number of ways you can create subsets of your n accessories for the next t days such that no accessory is worn two days in a row:

$$\sum_{x_1 \geq 0} \sum_{x_2 \geq 0} \cdots \sum_{x_t \geq 0} \binom{n}{x_1} \binom{n-x_1}{x_2} \binom{n-x_2}{x_3} \cdots \binom{n-x_{t-1}}{x_t} = F_{t+2}^n.$$

8. Stirling's Approximation

In this question, suppose $n \in \mathbb{Z}^+$, we want to find approximations for $n!$. For the parts that are marked with **[VL]**, please complete your answer in the Virtual Lab skeleton. You can also use an online tool (*e.g.*, go to http://www.wolframalpha.com/ and type "plot $\ln x$") if you wish to.

(a) **[VL]** Plot the function $f(x) = \ln x$.

(b) For the following three questions, please note that $\ln x$ is strictly increasing and concave-$\cap$ because, when $x > 0$, its first and second derivatives are positive and negative, respectively. Concavity means that all line segments connecting two points on the function are below the function.
Suppose $n \in \mathbb{Z}^+$, use the plot to explain why

$$\ln 1 + \ln 2 + \ldots + \ln n \geq \int_1^n \ln x dx \tag{1}$$

(c) Suppose $n \in \mathbb{Z}^+$, use the plot to explain why

$$\ln 1 + \ln 2 + \ldots + \ln n < \int_1^{n+1} \ln x dx \tag{2}$$

(d) Suppose $a \in \mathbb{Z}^+$, use the plot to explain why

$$\left(\frac{\ln a + \ln(a+1)}{2}\right) < \int_a^{a+1} \ln x dx \tag{3}$$

(e) Use Equation ((3)) to prove $n! \geq e\left(\frac{n}{e}\right)^n$.

(f) Use Equation ((4)) to prove $n! \leq en\left(\frac{n}{e}\right)^n$ (Hint: If in this part you find yourself wishing you had $n-1!$ on the left-hand-side, try to prove an upper bound on $n-1!$ and use that to help you)

(g) Use Equation ((5)) to prove $n! \leq e\sqrt{n}\left(\frac{n}{e}\right)^n$, which is a tighter upper bound.

(h) **[VL]** The Stirling's approximation is usually written as $n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$ or a simpler version $n! \approx \left(\frac{n}{e}\right)^n$. Plot the function $f(n) = \frac{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n}{n!}$. What do you observe?

(i) Suppose $m = \frac{k}{n}$, use m, n and apply the simpler version of the Stirling's approximation to rewrite $\binom{n}{k}$.

(j) **[VL]** Now, suppose $m_1 = \frac{k_1}{n} = 0.25$, $m_2 = \frac{k_2}{n} = 0.5$, and $m_3 = \frac{k_3}{n} = 0.75$, plot $\ln(\binom{n}{k_1})$, $\ln(\binom{n}{k_2})$, and $\ln(\binom{n}{k_3})$ as functions of n on a plot with linear-scaled axes. What do you observe?

9. Write your own problem

Write your own problem related to this week's material and solve it. You may still work in groups to brainstorm problems, but each student should submit a unique problem. What is the problem? How to formulate it? How to solve it? What is the solution?