EECS 70 Discrete Mathematics and Probability Theory
Spring 2014 Anant Sahai Homework 9

This homework is due Mar 31 2014, at 12:00 noon.

1. Virtual Lab 3: Biased Coins Continued

In this problem we will continue the lab from last HW. We will start from the optional problem at the end.

a) Up till this point, everything that you have done in this virtual lab 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.

So here, we will simply provide you with the right curve.

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 the earlier parts. (Do this integral numerically or use the special functions that most numerical packages already have for computing these sorts of integrals. This integral is related to something called the Error Function.) Notice how beautifully it hits the exact shape.

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

b) Just a little notation. We will use X_i to denote a random variable that is 1 if you toss a head on the i-th toss, and 0 if you toss a tail on the i-th toss. (This is just so we can more easily express things) So a sequence of k coin tosses would be $X_1, X_2, \ldots, X_{k-1}, X_k$, with each X_i being 0 or 1 depending on how the run actually came out. How would you write the total number S of heads as a function of the X_i's?

Our experience from the last homework tells us that the total number of heads is itself a random quantity since it varies based on the vagaries of the coin tosses.

c) Now, since you had realized earlier that the cliff-faces and the histograms have some natural relationship with each other, see if you can figure out a way to naturally overlay a smooth plot of $\frac{1}{2\pi} e^{-\frac{x^2}{2}}$ to the normalized histograms. What does this mean?

d) The other interesting pattern that you had seen in the previous virtual lab (In particular, in part k of Q1 on HW7) 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 $p = 0.3, 0.5, 0.7$ and a like $p + 0.05$, $p + 0.1$, $p + 0.2$. Take $m = 10000$ and plot the natural log of the frequencies these deviations against k (ranging from 10 to 200). Approximately extract the slopes for all 9 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-Liebler divergence and is also called the relative entropy.

Finally, add $e^{-D(a||p)k}$ to the plots (there should be 9 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.

Comment.

e) In the previous part, we went directly to one of the most powerful bounds we have. A much simpler bound (that we will rigorously derive later in the course) is called Chebyshev's inequality. For coin tosses, this inequality says

$$P(|S_k - kp| \geq \varepsilon k) \leq \frac{p(1-p)}{k\varepsilon^2}$$

.

Notice here that Chebyshev's inequality looks at two-sided deviations. We count both when S_k is much bigger than kp and when it is much smaller than kp. This is the difference from the previous bounds.

We can try to use this for the same kind of big deviations that we had examined above by trying $\varepsilon = 0.1, 0.2, 0.3$. Use simulations to compare what the actual frequency of such deviations is to what Chebyshev's inequality estimates.

Try to make an appropriate plot that shows both Chebyshev's inequality's prediction and where the actual frequencies are? Why is this hard?

f) Lastly, we would like to explore a function that defines how many ways you can choose k distinct objects out of n possible objects. This is written as $\binom{n}{k}$ and is read aloud as n choose k. We define $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ So, for example, $\binom{5}{3} = \frac{5!}{3!(5-3)!} = \frac{5\cdot4\cdot3\cdot2\cdot1}{(3\cdot2\cdot1)\cdot(2\cdot1)} = \frac{120}{12} = 10$. We wish to explore this function. Plot the value $\binom{10}{k}$ one the y-axis and k on the x-axis for $1 \leq k \leq n$. Does this constantly grow as k gets larger? Let's try plotting this for $n = 4$, $n = 10$, $n = 25$, $n = 50$. You should note a defining feature of the graph.

2. I'm Hungry... How many Pizzas?

Suppose that a pizzeria lets me choose toppings for the pizza among 10 different ingredients. I have budget for buying a pizza with 3 toppings. For this HW, you may express your answers as expressions, rather than as numbers (for example, $\frac{3!}{5!}$ instead of 0.05).

a) I choose 3 distinct toppings. If the order in which I order the toppings matters, (for example: ordering Tomatoes, Onions, Mushrooms is different than ordering Tomatoes, Mushrooms, Onions) how many ways are there to order the toppings?

b) Now I realize that it doesn't matter what order I ask for the toppings, I still get the same pizza. Again using 3 distinct toppings out of a possible 10, how many different types of Pizza can I order?

c) I don't have to get 3 toppings. If I can instead choose 0, 1, 2, or 3 distinct toppings. For example I could order tomatoes and onions and no third topping. How many different pizzas can I order now?

d) The waiter informs me that I can order repeat toppings (put 2 or 3 portions of the same topping). So now I can order Tomatoes, tomatoes, and onions, for example. Still sticking with my same budget, how many different pizzas can I order, making sure I order exactly 3 toppings total (although the toppings may not be distinct).

e) Under the same framework as above, how many pizzas can I order, this time I don't have to order exactly 3 toppings.

f) I go to a cheaper pizzeria, I can now order up to 5 toppings. How many pizzas can I buy? (Using the same rules as e above).

3. Hunger Games

In this modified version of the book (and movie), Esther is a young adult eligible for tribute to the Hunger Games. Esther is asked to pick 5 cards from a total of 100. Out of these 100 cards, 10 are marked. If at least 1 of the cards out of the 5 Esther picked is a marked card, Esther will go on as tribute to the Hunger Games.

a) How many ways in total are there for Esther to pick 5 out of 100 cards? The order in which she picks these cards does not matter.

b) How many ways can Esther choose the 5 cards such that all are unmarked? Now what is the probability that she will not go onto the Hunger Games?

c) What is the probability that she chooses 5 cards such that 1 is marked and the rest is unmarked? How about the configuration where 2 cards are marked and the rest is unmarked?

d) What is the probability that Esther will go onto the Hunger Games?

e) Given that Esther picked 1 marked card and 4 unmarked cards (she is going to the Hunger Games), her friend Tommy is now forced to choose 5 cards from the remaining pile. Calculate the probability that he will go to the Hunger Games.

4. Clinical Trials

You are creating a test for a rare disease that only 1 in 1000 people have. If a person has the disease, there is a 95% chance that your test will be positive. But if they don't have the disease, there is only an 85% chance the test will be negative.

a) Let D be the event that you have the disease, and H be the event that you are healthy. Let A be the event that the test comes out positive, and B be the event that it comes out negative. Write an expression for $\mathbb{P}(D|A)$ (the probability you have the disease given a positive test result) in terms of the probabilities given above. Evaluate your expression.

b) Write an expression for $P(H|B)$, the probability you are healthy given a negative test result.

c) For your test to gain approval, the chance of disease given a positive test result must be above 90%. What would the accuracy of the test have to be to ensure this result? (You may now assume the accuracy of the test is the same whether you have the disease or not.)

5. 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?