EECS 70 Discrete Mathematics and Probability Theory
Fall 2014 Anant Sahai

Homework 9

This homework is due November 3, 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. Intro to Randomness Lab (cont.)

In this week's lab, we will continue our coin tossing example, but see it from a different perspective. Make sure you review the lab solution from Homework 8 before moving on.

From this Virtual Lab onward, students who want to can also choose to completely rewrite the question. Basically, for parts (a) through (g) in this lab, read them, and then if you want, 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. Similarly for part (h), but this is its own separate thing.

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

(a) Let's change gears a little bit from last time. Consider the following visualization of a sequence of coin flips. We start at zero. For every head we get, we add one. For every tail we get, we subtract one. Hence, a sequence of 1000 coin tosses would be a path that starts at (0,0), and then goes to either $(1,1)$ or $(1,-1)$, and continues wandering till $(1000,y)$ somewhere. Plot 20 such paths on the same plot based on randomly flipped coins. Each sample path should have 1000 coin tosses.

What do you observe about the paths?

Hint: First, implement the `rand_one` function, which generates -1 and 1 randomly with roughly 50% probability each.

Then, implement the `path(n)` function, which returns a list of n elements that starts at 0 and every element thereafter is either one more or one less than the previous one. In Python, you can access the last element in a list with the syntax `lst[-1]`.

(b) Notice that the histograms you were plotting earlier were effectively looking at vertical slices in this picture and asking how many sample paths were crossing through a particular y coordinate. (If we are looking at k tosses, then having exactly h heads is the same as this sample path crossing through $(k,h-(k-h)) = (k,2h-k)$)

Now, let's see what the rescalings we were doing correspond to. The common-set-of-units scaling is what the previous part corresponded to. How would you change the scaling to correspond to the

normalized set of units in part (e) of last week's lab? (in this plot, a sample path that consists of all heads should basically be a straight line that stays at the upper-limit — say 1. And a sample path that consists of all tails should be a straight line that stays at the lower limit – say -1).

Give this new scaling (it will depend on k — so it will change the visual shape of a path) and plot 100 sample paths of 1000 coin tosses each.

Comment on what this suggests relative to the earlier plots.

(c) Shifting gears one more time, we are now going to look at the same basic experiment — tossing a fair coin k times — in a third way. Let R for a given run be the ratio of heads.

Fix $k = 1000$ to be the number of coin tosses in a run. Let $m = 1000$ be the number of runs. Plot how often $R \leq q$ as a function of q. The vertical axis should be (in linear scale) the fraction of the m runs in which $R \leq q$, while the horizontal scale should have q ranging from 0 to 1.

What do you notice about this curve?

Hint: Implement the function `q_curve`, which returns the sorted fraction of heads for each of the m runs. You may find Python's built-in `sorted` function helpful here.

(d) Repeat the previous part for different values of k and put them all on the same plot.

Try $k = 2, 10, 50, 100, 500, 10000$. (Recall that k is the number of tosses.) What do you see? Is this consonant with what you had observed in earlier plots?

(e) Now, think about rescaling the plots in the previous parts to see if there is something common about this shape. For each k, read off the q values where the curves seem to cross horizontal lines at $0.25, 0.5, 0.75$. Call these the quartile markers. Compute these qs for your experiment. Plot them as a function of k. What do you observe?

(f) Notice that the gap between the 0.75 marker and the 0.25 marker is getting smaller as k gets larger. Notice also that the 0.5 marker seems to be sticking around $q = \frac{1}{2}$. As a scientific problem, suppose you wanted to discover how indeed this was scaling with k.

Plot the distance between the 0.25 and 0.75 marker as a function of k as a scatter plot. Try all of the traditional axes combinations: linear-linear, log-linear, linear-log, and log-log. Which one seems to offer some insight?

Hint: In Python, the plotting functions with the aforementioned axes combinations are `plt.plot`, `plt.semilogx`, `plt.semilogy`, and `plt.loglog`.

(g) Based on what you observed in the previous set of plots, conjecture a scaling rule that lets you calculate the gap between the 0.75 marker and the 0.25 marker as a function of k for the fair coin tosses case. Explain your derivation in your writeup.

Use this rule to rescale the horizontal axis of the plots from three parts ago. What do you now observe about the curves for different values of k? By construction, they should be very close to each other in terms of where they are crossing $0.25, 0.5, 0.75$, but what about elsewhere?

Hint: Implement the function `q_curve_norm`, which does the same as `q_curve`, except now every point is normalized using your scaling rule.

Think about the previous part and how it can help you come up with a scaling rule.

(h) In the skeleton, you will find a simple implementation that simulates the Monty Hall problem. There are n doors, and only one contains the prize. The contestant first picks a door, and then the host Monty will open all $n-2$ doors that don't contain the prize. The contestant is then given a choice to switch or stay with his current choice.

Your task will be to simulate 10000 trials of the Monty Hall problem. What is the probability of winning when the contestant switches? How about when he/she stays? Does it match your expectation and what was described in lecture and the lecture note? Please report the result in your answer.

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 `hw9.zip`.

3. To Be Fair

Suppose you have a biased coin with $P(\textit{heads}) \neq 0.5$. How could you use this coin to simulate a fair coin? (Hint: Think about pairs of tosses.)

4. Intuition vs. Logic

(a) I have a bag containing either a \$5 bill (with probability $1/3$) or a \$10 bill (with probability $2/3$). I then add a \$5 bill to the bag, so it now contains two bills. The bag is shaken, and you randomly draw a bill from the bag without looking into the bag. Suppose it turns out to be a \$5 bill. If a second student draws the remaining bill from the bag, what is the probability that it, too, is a \$5 bill? Show your calculations.

(b) Your gambling buddy found a website where he could buy trick coins that are either heads or tails on both sides. He puts three coins into a bag: one coin that is heads on both sides, one coin that is tails on both sides, and one that is heads on one side and tails on the other side. You shake the bag, draw out a coin at random, put it on the table without looking at it, then look at the side that is showing. Suppose you notice that the side that is showing is heads. What is the probability that the other side is heads? Show your calculations.

5. Playing Pollster

As an expert in probability, the staff members at the Daily Californian have recruited you to help them conduct a poll to determine the percentage p of Berkeley undergraduates that plan to participate in the student sit-in. They've specified that they want your estimate $\hat{p}$ to have an error of at most ε with confidence $1-\delta$. That is,

$$P(|\hat{p}-p| \leq \varepsilon) \geq 1-\delta.$$

Assume that you've been given the bound

$$P(|\hat{p}-p| \geq \varepsilon) \leq \frac{1}{4n\varepsilon^2},$$

where n is the number of students in your poll.

(a) Using the formula above, what is the smallest number of students n that you need to poll so that your poll has an error of at most ε with confidence $1-\delta$?

(b) At Berkeley, there are about 26,000 undergraduates and about $10,000$ graduate students. Suppose you only want to understand the frequency of sitting-in for the undergraduates. If you want to obtain an estimate with error of at most 5% with 98% confidence, how many undergraduate students would you need to poll? Does your answer change if you instead only want to understand the frequency of sitting-in for the graduate students?

(c) It turns out you just don't have as much time for extracurricular activities as you thought you would this semester. The writers at the Daily Californian insist that your poll results are reported with at least 95% confidence, but you only have enough time to poll 500 students. Based on the bound above, what is the worst-case error with which you can report your results?

6. Blood Type

Consider the three alleles, A, B, and O, for human blood types. As each person inherits one of the 3 alleles from each parent, there are 6 possible genotypes: AA, AB, AO, BB, BO, and OO. Blood groups A and B are dominant to O. Therefore, people with AA or AO have type A blood. Similarly, BB and BO result in type B blood. The AB genotype is called type AB blood, and the OO genotype is called type O blood. Each parent contributes one allele randomly. Now, suppose that the frequencies of the A, B, and O alleles are 0.4, 0.25, and 0.35, respectively, in Berkeley. Alice and Bob, two residents of Berkeley are married and have a daughter, Mary. Alice has blood type AB.

(a) What is the probability that Bob's genotype is AO?

(b) Assume that Bob's genotype is AO. What is the probability that Mary's blood type is AB?

(c) Assume Mary's blood type is AB. What is the probability that Bob's genotype is AA?

7. Midterm question 3

Re-do midterm question 3.

8. Midterm question 4

Re-do midterm question 4.

9. Midterm question 5

Re-do midterm question 5.

10. Midterm question 6

Re-do midterm question 6.

11. Midterm question 7

Re-do midterm question 7.

12. Midterm question 8

Re-do midterm question 8.

13. Midterm question 9

Re-do midterm question 9.

14. Midterm question 10

Re-do midterm question 10.

15. Midterm question 11

Re-do midterm question 11.

16. Midterm question 12

Re-do midterm question 12.

17. Midterm question 13

Re-do midterm question 13.

18. Midterm question 14

Re-do midterm question 14.

19. 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?