Chris Piech
CS109

Oct 16, 2017

Problem Set #3

Due: ~~3pm~~ 11:59pm on Wednesday, Oct 25th

For each problem, explain/justify how you obtained your answer in order to obtain full credit. In fact, most of the credit for each problem will be given for the derivation/model used as opposed to the final answer. Make sure to describe the distribution and parameter values you used (e.g., Bin(10, 0.3)), where appropriate. Provide a numeric answer for all questions when possible.

Warmup

1. Recall the game set-up in the "St. Petersburg's paradox" discussed in class: there is a fair coin which comes up "heads" with probability $p = 0.5$. The coin is flipped repeatedly until the first "tails" appears. Let N = number of coin flips before the first "tails" appears (i.e., N = the number of consecutive "heads" that appear). Given that no one really has infinite money to offer as payoff for the game, consider a variant of the game where you win MIN($\$2^N$, X), where X is the maximum amount that the game provider will pay you after playing. Compute the expected payoff of the game for the following values of X. Show your work.

   a. X = $5.
   b. X = $500.
   c. X = $4096.

2. Lyft line gets 2 requests per 5 mins, on average, for a particular route. A user requests the route and Lyft commits a car to take her. All users who request the route in the next five minutes will be added to the car—as long as the car has space. The car can fit up to three users. Lyft will make $6 for each user in the car (the revenue) minus $7 (the operating cost). How much does Lyft expect to make from this trip?

3. Given our recent analysis of Justice Breyer's probabilistic arguments regarding jury selection, let's consider a situation involving juries. Suppose it takes at least 9 votes from a 12-member jury to convict a defendant. Suppose also that the probability that a juror votes that an actually guilty person is innocent is 0.2, whereas the probability that the juror votes that an actually innocent person is guilty is 0.1. If each juror acts independently and if 75% of defendants are actually guilty, find the probability that the jury renders a correct decision. Also determine the percentage of defendants found guilty by the jury.

4. The number of times a person's computer crashes in a month is a Poisson random variable with $\lambda = 5$. Suppose that a new operating system patch is released that reduces the Poisson parameter to $\lambda = 3$ for 75% of computers, and for the other 25% of computers the patch has no effect on the rate of crashes. If a person installs the patch, and has his/her computer crash 2 times in the month thereafter, how likely is it that the patch has had an effect on the user's computer (i.e., it is one of the 75% of computers that the patch reduces crashes on)?

5. Say there are k buckets in a hash table. Each new string added to the table is hashed to bucket i with probability p_i, where $\sum_{i=1}^{k} p_i = 1$. If n strings are hashed into the table, find the expected number of buckets that have at least one string hashed to them. (Hint: Let X_i be a binary variable that has the value 1 when there is at least one string hashed to bucket i after the n strings are added to the table (and 0 otherwise). Compute $E\left[\sum_{i=1}^{k} X_i\right]$.)

6. Say we have a cable of length n. We select a point (chosen uniformly randomly) along the cable, at which we cut the cable into two pieces. What is the probability that the shorter of the two pieces of the cable is less than 1/4th the size of the longer of the two pieces? Explain formally how you derived your answer.

7. Let X be a continuous random variable with probability density function:

$$f(x) = \begin{cases} c\,(3 - 2x^2) & -1 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

   a. What is the value of c?
   b. What is the cumulative distribution function (CDF) of X?
   c. What is E[X]?

Dithering.

8. [Coding] Two pseudo random number generators are used to simulate a sequence 300 independent flips of a fair coin (T means a tails was flipped, H means a head was flipped). Bellow are the two sequences (from the two random generators). Which one is a better random generator? Make an argument that is justified with probabilities calculated on the sequences:

   Sequence 1:
   TTHHTHTTHTTTHTTTHTTTHTTHTHHTHHTHTHHTTTHHTHTHTTHTHHTTHTHHTHTTTH
   HTTHHTTHHHTHHTHTTHTHTTHHTHHHTTHTHTTTHHTTHTHTHTHTHTTHTHTHHHTTHT
   HTHHTHHHTHTHTTHTTHHTHTHTHTTHHTTHTHTTHHHTHTHTHTTHTTHHTTHTHHTHHH
   TTHHTHTTHTHTHTHTHTHTHHHTHTHTHTTHTHHTHTHTTHTTTHHTHTTTHTHHTHHHT
   TTHHTHTHTHTHHHTTHHTHTTTHTHHTHTHTHHTHTTHTTHTHHTHTHTTT

   Sequence 2:
   HTHHHTHTTHHTTTTTTTTHHHTTTHHTTTTHHTTHHHTTHTHTTTTTTHTHTTTTHHHHTH
   THTTHTTTHTTHTTTTTHTHHTHHHHTTTTTHHHHTHHHTTTTHTHTTHHHHTHHHHHHHTT
   HHTHHTHHHHHHHTTHTHTTTHHTTTTHTHHTTHTTHTHTHTTHHHHHTTHTTTHTHTHHTT
   TTHTTTTTHHTHTHHHHTTTTHTHHHHHHTHTHTHTHHHTHTTHHHTHHHHHHTHHHTHTTT
   HHHTTTHHTHTTHHTHHHTHTTHTTHTTTTHHTHTHTTTTHTHTHTTHTHTHT

   The sequences are provided in the datasets zip as two files ditherSequence1.txt and ditherSequence2.txt.

Algorithmic Analysis

9. A bloom filter is a space-efficient, probabilistic set. In this problem we are going to look at it theoretically. Our bloom filter uses 3 different independent hash functions H_1, H_2, H_3 that each take any string as input and each return an index into a bit-array of length n. Each index is equally likely for each hash function. To add a string into the set, feed it to each of the 3 hash functions to get 3 array positions. Set the bits at all these positions to 1.

   For example, consider this bit-array of length $n = 10$. Values in the bit-array are initially zero:

Index:	0	1	2	3	4	5	6	7	8	9
Value:	0	0	0	0	0	0	0	0	0	0

   After adding a string "pie" where H_1("pie") = 4, H_2("pie") = 7 and H_3("pie") = 8, the bits at index {4, 7, 8} are set to be 1:

Index:	0	1	2	3	4	5	6	7	8	9
Value:	0	0	0	0	**1**	0	0	**1**	**1**	0

   Bits are never switched back to 0. Now, $m = 24{,}000$ strings are added to the bloom filter.

   a. Let $n = 8{,}000$. What is the (Poisson approximated) probability that the first bucket has 0 strings hashed to it?

   b. Let $n = 8{,}000$. What is the (Poisson approximated) probability that the first bucket has 10 or fewer strings hashed to it?

   To *check* whether a string is in the set, feed it to each of the 3 hash functions to get 3 array positions. If any of the bits at these positions is 0, the element is not in the set. If all bits at these positions are 1 the string is reported as in the set (though it might never have been added).

   c. Let $n = 100{,}000$. After $m = 25{,}000$ strings have been added to the bloom filter, what is the probability that a string, that has not been added to the set, will (incorrectly) be reported as in the set? Use approximations where appropriate.

   d. Our bloom filter uses three hash functions. Was that necessary? Repeat your calculation in (c) assuming that we only used a single hash function (not 3).

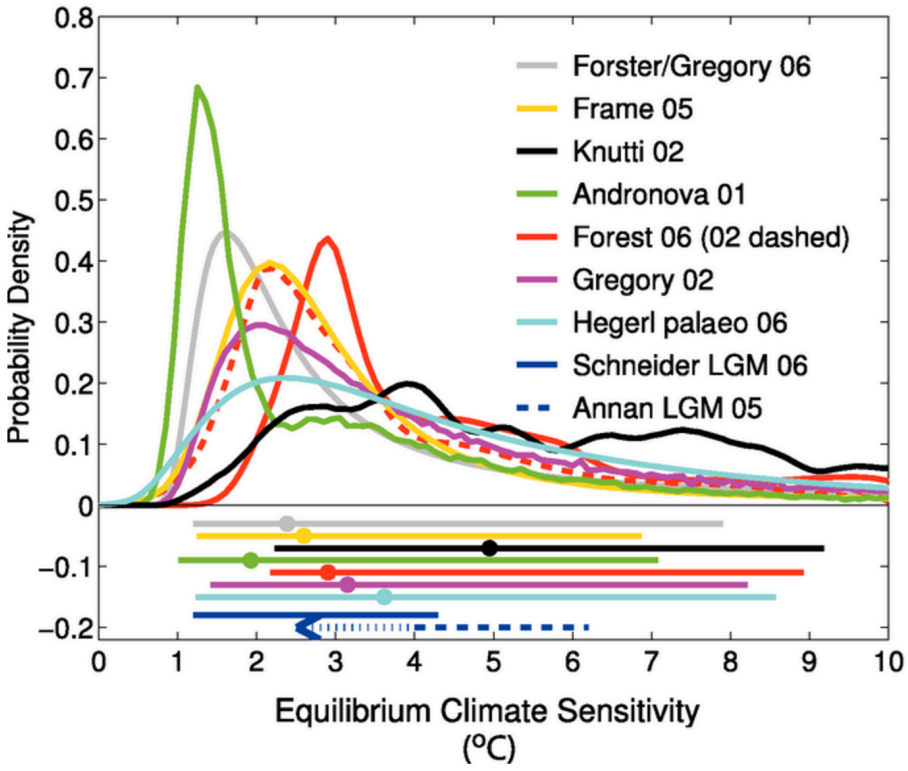
   (Chrome uses a Bloom Filter to keep track of malicious URLs. Questions such as this allow us to compute appropriate sizes for hash tables in order to get good performance with high probability in applications where we have a ballpark idea of the number of elements that will be hashed into the table).

Global Warming

10. On the day that this problem was released (Oct 16th, 2017) the concentration of CO_2 in the atmosphere was 403 parts per million (ppm) which is substantially higher than the pre-industrial concentration: 275 ppm. CO_2 is a greenhouse gas and as such increased CO_2 corresponds to a warmer planet.

    Absent some pretty significant policy changes we will reach a point within the next 50 years (eg well within your lifetime) where the CO_2 in the atmosphere will be double the pre-industrial level. In this problem we are going to explore the question: what will happen to the global temperature if atmospheric CO_2 doubles?

    The measure, in degrees Celsius, of how much the global average surface temperature will change (at the point of equilibrium) after a doubling of atmospheric CO_2 is called "Climate Sensitivity." Since the earth is a complicated ecosystem climate scientists model S as a random variable. The IPPC Fourth Assessment Report had a summary of 10 scientific studies that estimated the PDF for Climate Sensitivity (S):



    In this problem we are going to treat S as part-discrete and part-continuous. For values of S less than 7.5, we are going to model sensitivity as a discrete random variable with PMF based on the average of estimates from the studies in the IPCC report. Here is the PMF for S in the range 0 through 7.5:

Sensitivity, S (degrees C)	0	1	2	3	4	5	6	7
Expert Probability	0.00	0.11	0.26	0.22	0.16	0.09	0.06	0.04

The IPCC fifth assessment report notes that there is a non-negligible chance of S being greater than 7.5 degrees but didn't go into detail about probabilities. In the paper "Fat-Tailed Uncertainty in the Economics of Catastrophic Climate Change" Martin Weitzman discusses how different models for the PDF of Climate Sensitivity (S) for large values of S have wildly different policy implications.

For values of S greater than 7.5 degrees Celsius, we are going to model S as a continuous random variable. Consider two different assumptions for S when it is greater than 7.5: a fat tailed distribution (f_1) and a thin tailed distribution (f_2):

$$f_1(x) = \frac{K}{x} \text{ s.t. } 7.5 < x < 30 \qquad f_2(x) = \frac{K}{x^3} \text{ s.t. } 7.5 < x < 30$$

For this problem assume that the probability that S is greater than 30 degrees Celsius is 0.

a. Estimate the probability that Climate Sensitivity is greater than 7.5 degrees Celsius.
b. Calculate the value of K for both f_1 and f_2.
c. It is estimated that if temperatures rise more than 10 degrees Celsius, the ice on Greenland will melt. Estimate the probability that S is greater than 10 under both the f_1 and f_2 assumptions.
d. Calculate the expectation of S under both the f_1 and f_2 assumptions.
e. Let $R = S^2$ be a crude approximation of the cost to society that results from S. Calculate E[R] under both the f_1 and f2 assumptions.

Notes: (1) Both f_1 and f_2 are "Power law distributions" which are continuous forms of the Zipf distribution we talked about in class. (2) As mentioned in class, calculating expectations for a variable that is part discrete and part continuous is as simple as: use the discrete formula for the discrete part and the continuous formula for the continuous part.

Predicting Elections

11. [Coding] On May 7th 2017 France held an election between two candidates (candidate A and candidate B) to be their next president. By May 2nd there were 10 polls which each asked voters if they intend to vote for candidate A or B—we would like to see how we could have predicted the election. For each of the 10 polls we report their sample size (N samples), how many people said they would vote for candidate A (A votes), how many people said they would vote for candidate B (B votes). Not all polls are created equal—many have a bias towards one candidate or the other. For each poll we also report a value "weight" which represents how accurate we believe the poll was (see polls.csv):

Poll	N samples	A votes	B votes	Weight
1	862	548	314	0.93
2	813	542	271	0.85
3	984	682	302	0.82
4	443	236	207	0.87
5	863	497	366	0.89
6	648	331	317	0.81
7	891	552	339	0.98
8	661	479	182	0.79
9	765	609	156	0.63
10	523	405	118	0.68

a. First, assume that each sample in each poll is an independent experiment of whether or not a random person in France would vote for candidate A. In other words, there is no difference between a vote for candidate A in poll 1 vs a vote for candidate A in vote 7. Calculate the probability that a random person in France votes for candidate A.

b. The population of France is 64,888,792. Assume each person votes for candidate A with the probability calculated in the part (a) and otherwise votes for candidate B. What is the probability that candidate A gets more than half of the votes? Report your answer to two decimal places and explain how you computed it.

12. Nate Silvers at fivethirtyeight pioneered an approach called the "poll of polls" for predicting elections. For both candidates we are going to have a random variable which represents their strength on election night: variables S_A and S_B for candidates A and B respectively (this is the same ideas as ELO scores). The probability that A wins is $P(S_A > S_B)$.

a. Calculate the parameters for the random variables S_A and S_B. Both S_A and S_B are defined to be normal with the following parameters:

$$S_A \sim \mathcal{N}\Big(\mu = \sum_i p_{A_i} \cdot \text{weight}_i,\ \sigma^2\Big) \qquad S_B \sim \mathcal{N}\Big(\mu = \sum_i p_{B_i} \cdot \text{weight}_i,\ \sigma^2\Big)$$

where p_{A_i} is the ratio of A votes to N samples in poll *i*, p_{B_i} is the ratio of B votes to N samples in poll *i*, weight_i is the weight of poll *i*, m_i is the N samples in poll *i* and:

$$\sigma = \frac{K}{\sqrt{\sum_i m_i}} \text{ s.t. } K = 350$$

b. Calculate $P(S_A > S_B)$ by simulating 100,000 fake elections. In each fake election draw a random sample for the strength of A from S_A and a random sample for the strength of B from S_B. If S_A is greater than S_B, candidate A wins. Else candidate B wins. Report your answer to two decimal places.

c. Which model, the one from 11(b) or the model from 12(b) seems more appropriate. Explain briefly why that might be the case. On election night candidate A wins. Was your prediction from part (b) "correct"? Explain, briefly.