Chris Piech
CS 109

Problem Set #3
Apr 22, 2019

Problem Set #3
Due: 10am on Wednesday, May 1st

With problems by Mehran Sahami and Chris Piech

For each problem, briefly explain/justify how you obtained your answer. 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. Understanding the *process* that leads to different random variables is a great way to gain familiarity for what they mean. For each random variable, write a function that simulates its generation process. Your function should return a number. The only probability function that you may use when coding your solution is `random()`: a function that returns a uniform random in the range [0, 1]. Submit your code, either python or psuedocode. We include a solution to (a):

   a. $X \sim \text{Ber}(p = 0.4)$
   1 or 0 to indicate whether or not an underlying event was “successful".

```
def simulateBernoulli(p = 0.4):
    if random() < p:
        return 1
    return 0
```

   b. $X \sim \text{Bin}(n = 20, p = 0.4)$
   The number of successes after 20 independent experiments.
   c. $X \sim \text{Geo}(p = 0.03)$
   The number of trials until the first success.
   d. $X \sim \text{NegBin}(k = 5, p = 0.03)$
   The number of trials until 5 successes.
   e. $X \sim \text{Poi}(\lambda = 3.1)$ *approximate*
   The number of events in a minute, where the historical rate is 3.1 events per min.
   hint: break the minute down into 60,000 ms events like we did in lecture.
   f. $X \sim \text{Exp}(\lambda = 3.1)$ *approximate*
   The amount of time until the next event, where the historical rate is 3.1 events per min.
   hint: again think of an event for each ms.

   If you are trying to understand probability mass functions, you may optionally try to visualize one via your simulations. Run one of your simulations 100,000 times and plot a histogram of return values.

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).
    a. How much does Lyft expect to make from this trip?
    b. Lyft has one space left in the car and wants to wait to get another user. What is the probability that another user will make a request in the next 15 seconds?

3. 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 0.25, whereas the probability that the juror votes that an actually innocent person is guilty is 0.15. If each juror acts independently and if 70% 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. 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 $E\left[\sum_{i=1}^{k} X_i\right]$.)

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

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

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

6. Scores on the SAT maths (out of 800) are normally distributed with a mean of 500 and a standard deviation of 100.
    a. What fraction of students receive a score within 1.5 standard deviations of the mean?
    b. Irina scores 750. What percent of students scored lower than 750? (Irina's percentile)

7. Let X be a Normal random variable with $\mu = 6$. If $P(X > 9) = 0.3$, what is the approximate value of $\text{Var}(X)$?

8. The Huffmeister floodplane in Houston has historically been estimated to flood at an average rate of 1 flood every 500 years. A flood plane with that rate of flooding is called a "500 year" floodplane.
    a. What is the probability of observing at least 3 floods in 500 years?
    b. What is the probability that a flood will occur within the next 100 years?
    c. What is the expected number of years until the next flood?

9. You are testing software and discover that your program has a non-deterministic bug that causes catastrophic failure (aka a "hindenbug"). Your program was tested for 400 hours and the bug occurred **twice**.

a. Each user uses your program to complete a three hour long task. If the hindenbug manifests they will immediately stop their work. What is the probability that the bug manifests for a given user?
b. Your program is used by one million users. Use a normal approximation to estimate the probability that more than 10,000 users experience the bug. Use your answer from part (a).

Dithering

10. 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:
    TTHHTHTTHTTTHTTTHTTTHTTHTHHTHHTHTHHTTTHHTHTHTTHTHHTTHTHHT
    HTTTHHTTHHTTHHHTHHTHTTHTHTTHHTHHHTTHTHTTTHHTTHTHTHTHTHTT
    HTHTHHHTTHTHTHHTHHHTHTHTTHTTHHTHTHTHTTHHTTHTHTTHHHTHTHTH
    TTHTTHHTTHTHHTHHHTTHHTHTTHTHTHTHTHTHTHHHTHTHTHTTHTHHTHTH
    TTHTTTHHTHTTTHTHHTHHHHTTTHHTHTHTHTHHHTTHHTHTTTHTHHTHTHTH
    HTHTTHTTHTHHTHTHTTT

    Sequence 2:
    HTHHHTHTTHHTTTTTTTTHHHTTTHHTTTTHHTTHHHTTHTHTTTTTTHTHTTTTH
    HHHTHTHTTHTTTHTTHTTTTHTHHTHHHHTTTTTHHHHTHHHTTTTHTHTTHHHH
    THHHHHHHHTTHHTHHTHHHHHHHTTHTHTTTHHTTTTHTHHTTHTTHTHTHTTHH
    HHHTTHTTTHTHTHHTTTTHTTTTTHHTHTHHHHTTTTHTHHHHHHTHTHTHTHHH
    THTTHHHTHHHHHHTHHHTHTTTHHHTTTHHTHTTHHTHHHTHTTHTTHTTTHHTH
    THTTTTHTHTHTTHTHTHT

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

Analysis of Bloom Filters

11. A Bloom filter is a probabilistic implementation of the *set* data structure, an unordered collection of unique objects. 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, initially all values in the bit-array are zero. In this example $n = 10$:

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:

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. Consider a Bloom filter with $n = 9,000$ buckets. You have added m = 1,000 strings to the Bloom filter. Provide a **numerical answer** for all questions.

    a. What is the (approximated) probability that the first bucket has 0 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 *may* be in the set; but it could be that those bits are 1 because some of the other strings hashed to the same values. You may assume that the value of one bucket is independent of the value of all others.

    c. What is the probability that a string which has *not* previously been added to the set will be misidentified as in the set. That is, what is the probability that the bits at all of its hash positions are already 1? Use approximations where appropriate.
    d. Our bloom filter uses three hash function. Was that necessary? Repeat your calculation in (c) assuming that we only use a single has function (not 3).

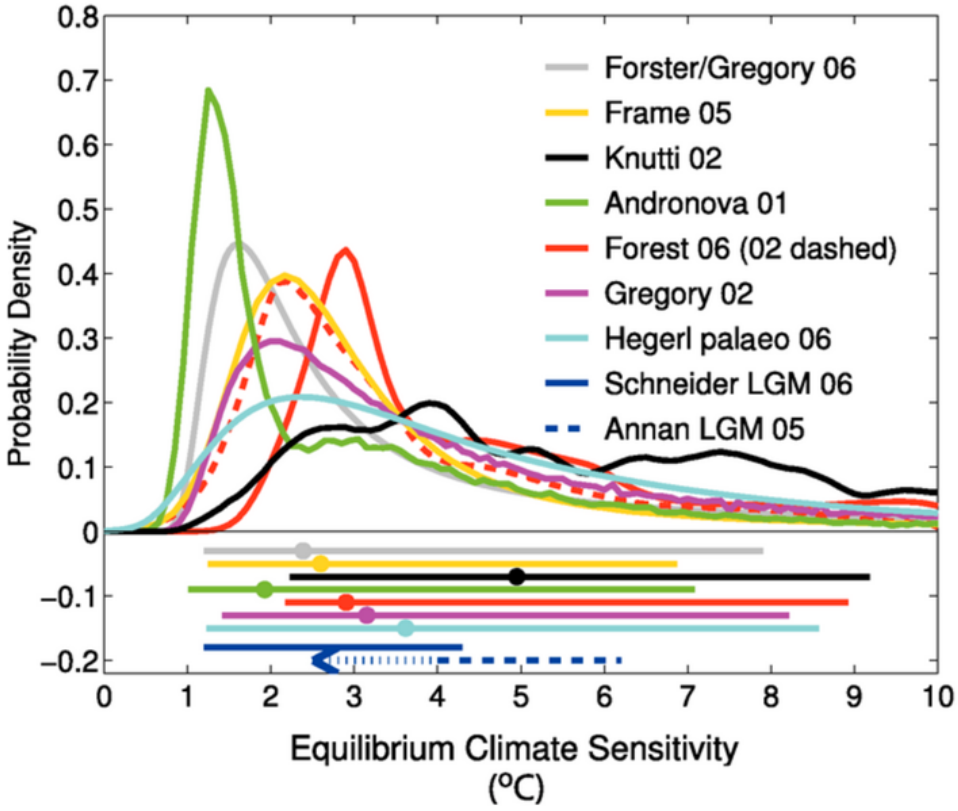
    (Chrome uses a Bloom filter to keep track of malicious URLs. Questions such as this allow us to compute appropriate sizes for hash tables in order to get good performance with high probability in applications where we have a ballpark idea of the number of elements that will be hashed into the table.)

Climate Change

12. This summer (Jun 2018) the concentration of CO2 in the atmosphere was 411 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.âĂİ 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):

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

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

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 f1 and f2.
c. It is estimated that if temperatures rise more than 10 degrees Celsius, all 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 f_2 assumptions.

Notes: (1) Both f1 and f2 are "power law distributions". (2) 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.