Jerry Cain
CS 109

Problem Set #3
April 19, 2021

Problem Set #3
Due: 10:00am on Friday, April 30th

With problems by Mehran Sahami, Chris Piech, and Lisa Yan

- Submit on Gradescope by 10:00am Pacific on Friday, April 30th, for a small, "on-time" bonus.
- All students have a pre-approved extension, or "grace period" that extends until Monday 10:00am Pacific, when they can submit with no penalty. **The grace period expires on 10:00am Pacific on Monday,, May 3rd**, after which we cannot accept any more submissions.
- **Collaboration policy:** You are encouraged to discuss problem-solving strategies with each other as well as the course staff, but you must write up your own solutions and submit individual work. Please cite any collaboration at the top of your submission.
- **Tagging written problems**: When you submit your written PDF on Gradescope you must tag your PDF, meaning that you must assign pages of your PDF as answers to particular questions so that we can properly grade your submission. For problem sets, we are deducting **2 points** for any submissions that do not have all questions tagged.
- **For each problem, briefly explain/justify how you obtained your answer.** Brief explanations of your answer are necessary to get full credit for a problem even if you have the correct numerical answer. The explanations help us determine your understanding of the problem whether or not you got the correct answer. Moreover, in the event of an incorrect answer, we can still try to give you partial credit based on the explanation you provide. It is fine for your answers to include summations, products, factorials, exponentials, or combinations; you don't need to calculate those all out to get a single numeric answer.
- If you handwrite your solutions, you are responsible for making sure that you can produce **clearly legible** scans of them for submission. You may also use any word processing software you like for writing up your solutions. On the CS109 webpage we provide a template file and tutorial for the LaTeX system, if you'd like to use it.

Coding Problem

Download the starter code for this problem from the Problem Set #3 webpage. Submit your completed file on Gradescope under "PSet 3 - Coding". We expect you to follow these guidelines:

- Do not use global variables.
- You may define helper functions if you wish.
- Your code should not print anything.

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 `numpy.random.rand()`: a function that returns a uniform random in the range [0, 1]. We include a solution to (a) in the starter code and below. Note that a function from one part may call a function from a previous part if you wish.

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

```python
from numpy.random import rand

def simulate_bernoulli(p=0.4):
    if rand() < 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. $X \sim \text{NegBin}(r = 5, p = 0.03)$
   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: Like part (e), think of an event for each millisecond.

Written Problems

Submit your solutions to these written problems as a single pdf file on Gradescope.

2. Ryde is a new nonprofit van share service where a driver gets paid all of the generated revenue, minus operating costs. A Ryde customer requests a route and the Ryde application commits a van to take them (i.e. there is already one person waiting in the vehicle). Any additional customers who request the same route in the next five minutes are added as long as it has space. Each Ryde van can fit up to five passengers, and on average, each Ryde driver gets 2 requests every 5 minutes for a particular route (and all requests are assumed to be independent). The Ryde driver will make \$12 for each passenger in the car (the per-passenger charge) minus \$3 (the operating cost).

   a. How much does the Ryde driver expect to make from any single trip?

   b. Assume the Ryde driver has one space left in the van and is willing to wait for another passenger. What is the probability another passenger will make a request in the next 30 seconds?

3. To determine whether they have measles, 1000 people have their blood tested. However, rather than testing each individual separately (1000 tests is quite costly), it is decided to use a *group testing* procedure:

   - Phase 1: First, place people into groups of 5. The blood samples of the 5 people in each group will be pooled and analyzed together. If the test is positive (at least one person in the pool has measles), continue to Phase 2. Otherwise send the group home. 200 of these pooled tests are performed.
   - Phase 2: Individually test each of the 5 people in the group. 5 of these individual tests are performed per group in Phase 2.

   Suppose that the probability that a person has measles is 5% for all people, independently of others, and that the test has a 100% true positive rate and 0% false positive rate (note that this is unrealistic). Using this strategy, compute the expected total number of blood tests (individual and pooled) that we will have to do across Phases 1 and 2.

4. The number of times a person's computer crashes in a month is a Poisson random variable with $\lambda = 7$. Suppose that a new operating system patch is released that reduces the Poisson parameter to $\lambda = 2$ for 80% of computers, and for the other 20% of computers the patch has no effect on the rate of crashes. If a person installs the patch, and has their computer crash 4 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 80% of computers that the patch reduces crashes on)?

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. Consider a hash table with n buckets. Now, m strings are hashed into the table (with equal probability of being hashed into any bucket).

   a. Let $n = 2{,}000$ and $m = 10{,}000$. What is the (Poisson approximated) probability that the first bucket has 0 strings hashed to it?
   b. Let $n = 2{,}000$ and $m = 10{,}000$. What is the (Poisson approximated) probability that the first bucket has 8 or fewer strings hashed to it?
   c. Let $m = 10{,}000$. What is largest integer value n such the Poisson approximated probability that an arbitrary bucket in the hash table will have no strings hashed to it is less than 0.5 (= 50%)?
   d. Let X be a Poisson random variable with parameter λ, that is: $X \sim \text{Poi}(\lambda)$. What value of λ maximizes $P(X = 3)$? Show formally (mathematically) how you derived this result. (Hint: at some point in your derivation you should be differentiating with respect to λ.)

   Questions like 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 to be stored.

7. Say the lifetimes of computer chips produced by a certain manufacturer are normally distributed with parameters $\mu = 1.5 \times 10^6$ hours and $\sigma = 9 \times 10^5$ hours. The lifetime of each chip is independent of the other chips produced.
   a. What is the **approximate** probability that a batch of 100 chips will contain at least 6 whose lifetimes are more than 3.0×10^6 hours?
   b. What is the **approximate** probability that a batch of 100 chips will contain at least 65 whose lifetimes are less than 1.9×10^6 hours? Provide a numeric answer for this part.

8. An election has two candidates in a very close race: recent polls predict that candidate A will win about 51% of the vote, while candidate B will win about 49%.
   a. Suppose there are $N = 5{,}000$ voters in the election, and that every voter in the election votes randomly and independently with those probabilities: 0.51 for candidate A, 0.49 for candidate B. Give an expression (involving a sum) for the probability that candidate A wins the election (gets strictly more than $N/2 = 2{,}500$ votes).
   b. Compute the numerical value of this probability. (Remember you can use `scipy.stats` or similar libraries in other languages. Present your Python code—which is optimally just one line—in your written submission, but don't look to upload it to be autograded.)
   c. Compute the numerical value of the same probability using a Poisson approximation. Is Poisson a good approximation here? Why or why not?
   d. Compute the numerical value of the same probability using a Normal approximation. Is the Normal a good approximation here? Why or why not?

9. 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 probability that the first bucket has 0 strings hashed to it? You can provide either an exact or approximate probability.

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.

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

Incidentally, 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.

10. Last summer (May 2019) the concentration of CO_2 in the atmosphere was 414 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 (i.e., 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 following question: What will happen to the global temperature if atmospheric CO_2 doubles?

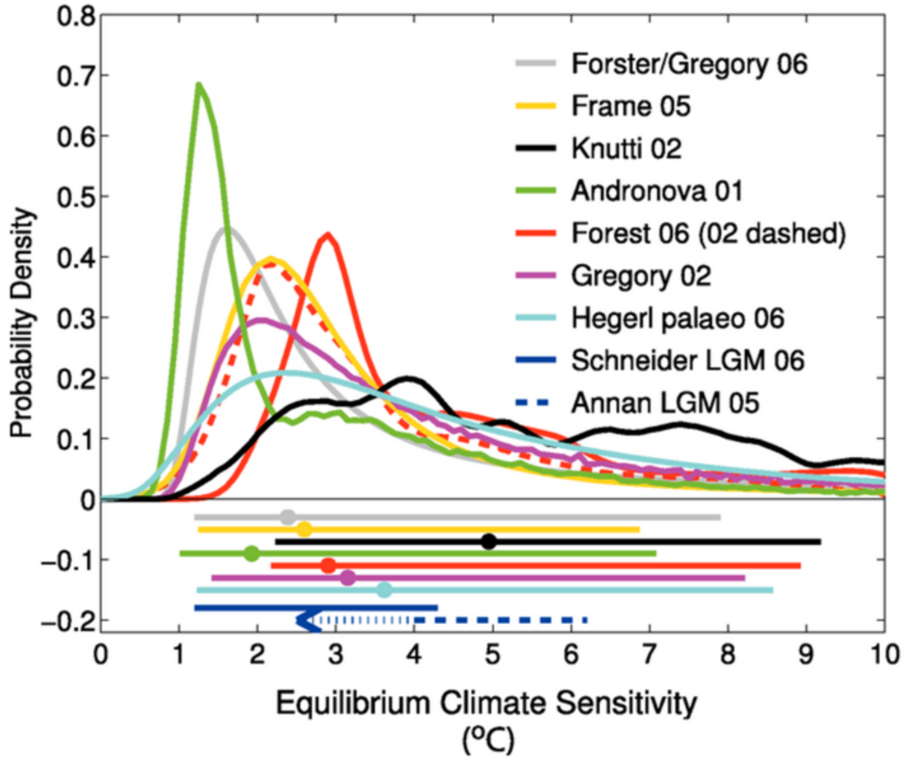
The measure, in degrees Celsius, of how much the global average surface temperature will change (at the point of equilibrium) after a doubling of atmospheric CO_2 is called "Climate Sensitivity." Since the earth is a complicated ecosystem climate scientists model Climate Sensitivity as a random variable, S. The IPPC Fourth Assessment Report had a summary of 10 scientific studies that estimated the PDF of 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 or equal to 7.5 degrees Celsius, we are going to model S as a continuous random variable. Consider two different assumptions for S when it is at least 7.5 degrees Celsius: a fat tailed distribution (f_1) and a thin tailed distribution (f_2):



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

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

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

a. Compute the probability that Climate Sensitivity is at least 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, 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 f_1 and f_2 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.