Chris Piech
CS 109

Problem Set #3
Oct 11, 2021

Problem Set #3
Due: 1:00pm on Friday, Oct 22nd

With problems by Mehran Sahami, Chris Piech

Submit on Gradescope by 1:00pm Pacific on Friday, Oct 22nd, for a small, "on-time" bonus. All students can give thmeselves a pre-approved extension, or "grace period" that extends until Monday 1:00pm Pacific, when they can submit with no penalty. **The grace period expires on 1:00 Pacific on Monday, Oct 25th**.

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. It is fine for your answers to include summations, products, factorials, exponentials, or combinations, unless you are specifically asked for a computed numerical answer.

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) below. Note that a function from one part may call a function from a previous part if you wish. Submit your code (or pseudocode).

   For extra credit, call your simulation function 1,000 times and include a histogram of how often each different values occurred. For part f discretize time into reasonable sized buckets.

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

```
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 intervals 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 interval for each millisecond.

2. The **mode** of a random variable is the value that it can take on with the *highest probability*. For example the mode of $X \sim \text{Bern}(0.9)$ is 1. For each of the following distributions calculate the mode and, for contrast, the expectation. To show your work, include the probability of the mode, as well as one integer lower and higher than the mode.

   a. $X \sim \text{Bin}(n = 18, p = 0.16)$
   b. $X \sim \text{Geo}(p = 0.2)$
   c. $X \sim \text{NegBin}(r = 3, p = 0.19)$

3. GrabShare is a ride-sharing service which is popular in South East Asia. GrabShare gets 2 requests every 5 minutes, on average, for a particular route. A user requests the route and GrabShare 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 passengers. The driver will make S\$6 for each user in the car (the revenue) minus S\$7 (their operating cost). Note that S\$ stands for Singaporean dollar.

   a. How much does driver expect to make from this trip?
   b. GrabShare has one space left in the car and wants to wait to get another passenger. What is the probability that another passenger will make a request in the next 30 seconds?

4. Suppose it takes at least 9 votes from a panel of 12 judges to win a competition. The probability that a judge votes for you is 0.85. If each judge acts independently, find the probability that you win.

5. To determine whether they have measles, 60 people have their blood tested. However, rather than testing each individual separately, it is decided to first place the people into groups of 6. The blood samples of the 6 people in each group will be pooled and analyzed together. If the test is negative, one test will suffice for the 6 people, whereas if the test is positive, each of the 6 people will also be individually tested and, in all, 7 tests will be made on this group. Note that we assume that the pooled test will be positive if at least one person in the pool has measles. Assume that the probability that a person has measles is 5% for all people, independently of each other, and compute the expected number of tests necessary for each group of 6 people.

6. 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 $E[X]$?

7. 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).

8. The **median** of a continuous random variable having cumulative distribution function F is the value m such that $F(m) = 0.5$. That is, a random variable is just as likely to be larger than its median as it is to be smaller. Find the median of X (in terms of the respective distribution parameters) in each case below.

   a. $X \sim \text{Uni}(a, b)$
   b. $X \sim \text{N}(\mu, \sigma^2)$
   c. $X \sim \text{Exp}(\lambda)$

9. 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?

   (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. 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.

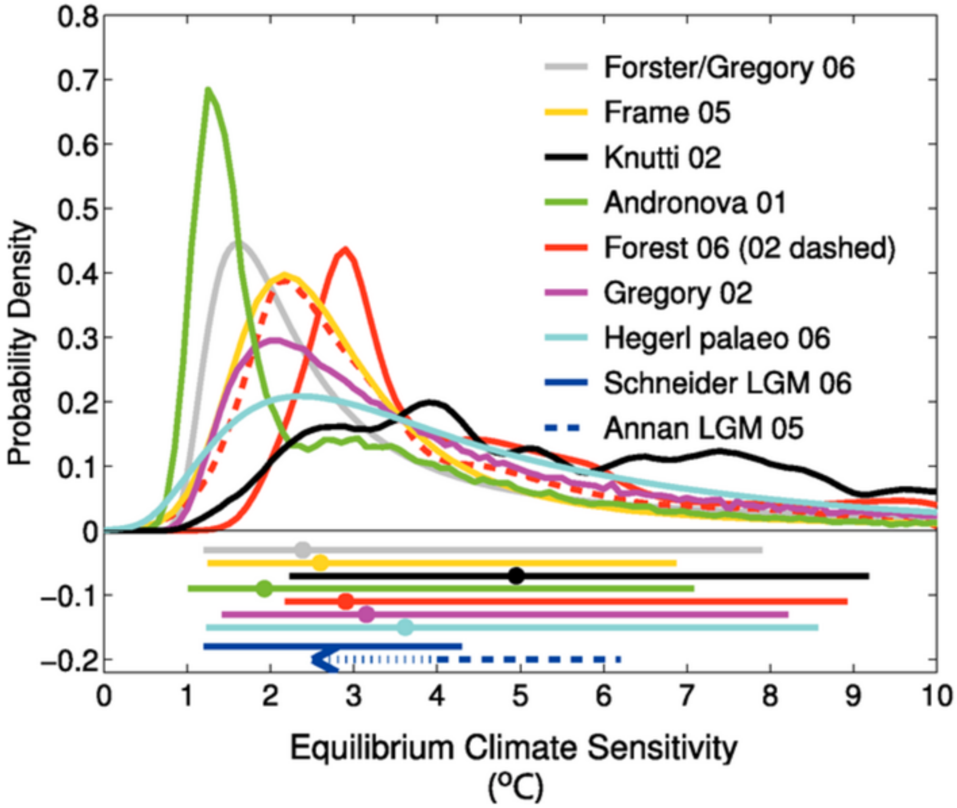
b. What is the probability that a string which has *not* previously been added to the set will be misidentified as in the set? That is, what is the probability that the bits at all of its hash positions are already 1? Use approximations where appropriate.
c. Our Bloom filter uses three hash functions.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

(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.)

11. Last summer (May 2021) the concentration of CO_2 in the atmosphere was 420 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 Fifth 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.

12. Extra Credit: Below are two sequences of 300 "coin flips" (H for heads, T for tails). One of these is a true sequence of 300 independent flips of a fair coin. The other was generated by a person typing out H's and T's and trying to *seem* random. Which sequence is truly composed of coin flips?

    We'll save you a bit of time by telling you that both sequences have 148 heads, two less than the expected number for a 0.5 probability of heads. It won't be as simple as finding out which one is closer to half heads! Make an argument that is justified with probabilities calculated on the sequences. This problem is solvable without code, but it would require some tedious counting. You're encouraged to put your computer to good use by looking at these sequences in the accompanying `pset3.zip`. The answer you submit should not include any code – rather it should be a short argument (under 500 words). You can optionally include figures.

    Sequence 1:

    ```
    TTHHTHTTHTTTHTTTHTTTHTTHTHHTHHTHTHHTTTHHTHTHTTHTHH
    TTHTHHTHTTTHHTTHHTTHHHTHHTHTTHTHTTHHTHHHTTHTHTTTHH
    TTHTHTHTHTHTTHTHTHHHTTHTHTHHTHHHTHTHTTHTTHHTHTHTHT
    THHTTHTHTTHHHTHTHTHTTHTTHHTTHTHHTHHHTTHHTHTTHTHTHT
    HTHTHTHHHTHTHTHTHHTHHTHTHTTHTTTHHTHTTTHTHHTHHHHTTT
    HHTHTHTHTHHHTTHHTHTTTHTHHTHTHTHHTHTTHTTHTHHTHTHTTT
    ```

    Sequence 2:

    ```
    HTHHHTHTTHHTTTTTTTTHHHTTTHHTTTTHHTTHHHTTHTHTTTTTTH
    THTTTTHHHHTHTHTTHTTTHTTHTTTTHTHHTHHHHTTTTTHHHHTHHH
    TTTTHTHTTHHHHTHHHHHHHHTTHHTHHTHHHHHHHTTHTHTTTHHTTT
    THTHHTTHTTHTHTHTTHHHHHTTHTTTHTHTHHTTTTHTTTTTHHTHTH
    HHHTTTTHTHHHTHHTHTHTHTHHHTHTTHHHTHHHHHHTHHHTHTTTHH
    HTTTHHTHTTHHTHHHTHTTHTTHTTTHHTHTHTTTTHTHTHTTHTHTHT
    ```