Chris Piech
CS109

PSet #4
May 3, 2017

Problem Set #4
Due: 10:30am on Monday, May 17th

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

Warmup

1. On average 5.5 users sign-up for an on-line social networking site each minute. What is the probability that:
   a. More than 7 users will sign-up for the social networking site in the next minute?
   b. More than 13 users will sign-up for the social networking site in the next 2 minutes?
   c. More than 15 users will sign-up for the social networking site in the next 3 minutes?

2. The joint probability density function of continuous random variables X and Y is given by:
   $f_{X,Y}(x,y) = c\frac{y}{x}$ where $0 < y < x < 1$
   a. What is the value of c in order for $f_{X,Y}(x,y)$ to be a valid probability density function?
   b. Are X and Y independent? Explain why or why not.
   c. What is the marginal density function of X?
   d. What is the marginal density function of Y?
   e. What is E[X]?
   f. What is E[Y]? Hint: At some point, integration by parts may be your friend on this problem. You may use Wolfram Alpha or a similar integration tool.

3. Let X_i = the number of weekly visitors to a web site in week i, where $X_i \sim N(2200, 52900)$ for all i. Assume that all X_i are independent of each other.
   a. What is the probability that the total number of visitors to the web site in the next two weeks exceeds 5000.
   b. What is the probability that the weekly number of visitors exceeds 2000 in at least 2 of the next 3 weeks?

4. Let X, Y, and Z be independent random variables, where $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$, and $Z \sim N(\mu_3, \sigma_3^2)$.
   a. Let A = X + Y. What is the distribution of A?
   b. Let B = 5X + 2. What is the distribution of B?

c. Let C = $aX - bY + c^2Z$, where a, b, and c are real-valued constants. What is the distribution (along with parameter values) for C? Show how you derived your answer.

Deeper Questions

5. Choose a number X at random from the set of numbers {1, 2, 3, 4, 5, 6}. Now choose a number at random from the subset no larger than X, that is from {1, …, X}. Let Y denote the second number chosen.

   a. Determine the joint probability mass function of X and Y.

   b. Determine the conditional mass function P($X = j \mid Y = i$) as a function of i and j.

   c. Are X and Y independent? Justify your answer.

6. A robot is located at the center of a square world that is 10 kilometers on each side. A package is dropped off in the robot's world at a point (x, y) that is uniformly (continuously) distributed in the square. If the robot's starting location is designated to be (0, 0) and the robot can only move up/down/left/right parallel to the sides of the square, the distance the robot must travel to get to the package at point (x, y) is |x| + |y|. Let D = the distance the robot travels to get to the package. Compute E[D]. *The distance calculation used in this problem is often called the "L1 Norm" and is a common metric for many problems.*

7. Say we have an array of n doubles, `arr[n]` (indexed from 0 to n – 1), which contains uniformly generated *non-negative* real values (where each value in the array is unique). What is the expected number of times that "max update" (as noted by the comment in the code) is executed in the function below (assuming the function is passed the array `arr` and its size `n`). Give an expression (not a big-Oh running time) for the expectation, and explain how you derived your answer.

```
double max(double arr[], int n) {
   double max = -1;        // note: all elements in arr[] are > -1.
   for(int i = 0; i < n; i++) {
      if (arr[i] > max) {
         max = arr[i];     // max update: (max = arr[i])
      }
   }
   return max;
}
```

8. Say we have a coin with unknown probability X of coming up heads when flipped. However, we believe (subjectively) that the prior probability (before seeing the results of any flips of the coin) of X is a Beta distribution, where E[X] = 0.5 and Var(X) = 1/36 ≈ 0.02778.

   a. What are the values of the parameters a and b (where a, $b > 1$) of the prior Beta distribution for X?

   b. Now say we flip the coin 13 times, obtaining 8 heads and 5 tails. What is the form (and parameters) of the posterior distribution of (X | 13 flips resulting in 8 heads and 5 tails)?

   c. What is E[X | 12 flips resulting in 8 heads and 4 tails]?

   d. What is Var(X | 12 flips resulting in 8 heads and 4 tails)?

Dithering.

9. 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
   THTTHTTTHTTHTTTTHTHHTHHHHTTTTTHHHHTHHHTTTTHTHTTHHHHTHHHHHHHTT
   HHTHHTHHHHHHHTTHTHTTTHHTTTTHTHHTTHTTHTHTHTTHHHHHTTHTTTHTHTHTT
   TTHTTTTTHHTHTHHHHTTTTHTHHHHHHTHTHTHTHHHTHTTHHHTHHHHHHTHHHTHTTT
   HHHTTTHHTHTTHHTHHHTHTTHTTHTTTHHTHTHTTTTHTHTHTTHTHTHT

Biometric Keystrokes.

10. Did you know that computers can know who you are, not just by what you write, but also by how you write it? Coursera uses Biometric Keystroke signatures for plagiarism detection. If you can't write a sentence with the same statistical distribution of key press timings as in your previous work, they assume that it is not you who is sitting behind the computer. In this problem we provide you with three files:

    personKeyTimingA.txt has keystroke timing information for a user A writing a passage. The first column is the time in milliseconds ***(since the start of writing)*** when the user hit each key. The second column is the key that the user hit.

    personKeyTimingB.txt has keystroke timing information for a second user (user B) writing the same passage as the user A. Even though the content of the passage is the same *the timing* of how the second user wrote the passage is different.

    email.txt has keystroke timing information for an unknown user. We would like to know if the author of the email was user A or user B.

    Let X and Y be random variables for the duration of time , in milliseconds since the last keystroke, for users A and B (respectively) to type a key. Assume that each keystroke from a user has a duration that is an independent random variable with the same distribution.

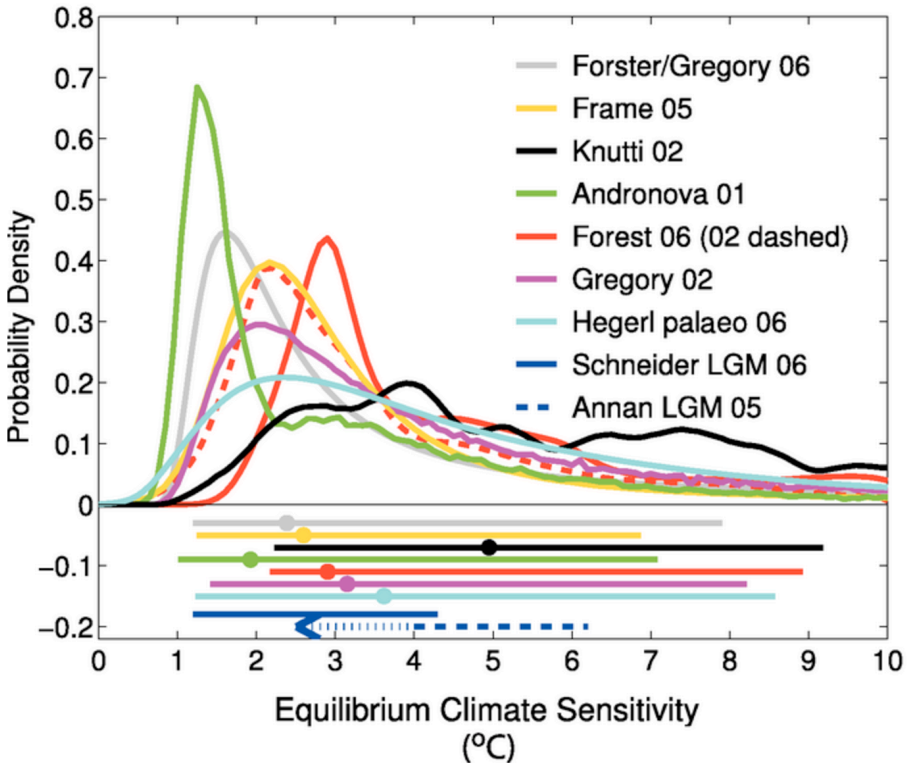
    a. Estimate E[X] and E[Y]
    b. Estimate $E[X^2]$ and $E[Y^2]$
    c. Use your answers to part (a) and (b) and approximate X and Y as Normals with mean and variance that match their biometric data. Report both distributions.
    d. Calculate the ratio of the probability that user A wrote the email over the probability that user B wrote the email.

Global Warming

11. On the day that this problem set was written (April 29th, 2017) the concentration of CO_2 in the atmosphere was 407 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 [2].

    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) [1]:



    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 [3].

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 f_2 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.

[1] IPCC. Climate Change 2007: Working Group I: The Physical Science Basis. https://www.ipcc.ch/publications_and_data/ar4/wg1/en/ch10s10-5.html
[2] IPCC. Climate Change 2014: Summary for Policymakers. https://www.ipcc.ch/pdf/assessment-report/ar5/syr/AR5_SYR_FINAL_SPM.pdf
[3] Weitzman, Martin. Fat-Tailed Uncertainty in the Economics of Catastrophic Climate Change. https://scholar.harvard.edu/files/weitzman/files/fattaileduncertaintyeconomics.pdf