Homework 6: Continuous Random Variables

For each problem, remember you must briefly explain/justify how you obtained your answer, as correct answers without an explanation will not receive full credit. Moreover, in the event of an incorrect answer, we can still try to give you partial credit based on the explanation you provide.

In general, your goal in an explanation is to write enough that a student from class who has attended lecture, but not read the problem yet, could understand your approach, verify your reasoning, and believe your answer is correct. While we do not usually need to see arithmetic, you must include enough work that in principle one could rederive your answer with only a scientific calculator.

Unless a problem states otherwise, you should leave your answer in terms of factorials, combinations, etc., for instance 26^7 or $26!/7!$ or $26 \cdot \binom{26}{7}$ are all good forms for final answers.

Instructions as to how to upload your solutions to gradescope are on the course web page.

Remember that you must tag your written problems on Gradescope.

Submission: You must upload a **pdf** of your written solutions to Gradescope under "HW 6 [Written]". (Instructions as to how to upload your solutions to gradescope are on the course web page.) The use of latex is *highly recommended*. (Note that if you want to hand-write your solutions, you'll need to scan them. We will take off points for hand-written solutions that are difficult to read due to poor handwriting and neatness.)

Your code for Problem 6 will be submitted under "HW 6 [Coding]" as a file called `min_hash.py`.

Due Date: This assignment is due at 11:59 PM Wednesday May 15.

Collaboration: Please read the full collaboration policy. If you work with others (and you should!), you must still write up your solution independently and name all of your collaborators somewhere on your assignment.

0. Extra Instructions :o

For calculations that require evaluating integrals (unless we indicate otherwise), you must

(a) Show the integral to evaluate (e.g., $\int_0^2 z \cdot 2dz$)

(b) Show an antiderivative and the values to evaluate at (e.g., $z^2|_0^2$)

(c) Plug in the values and simplify (e.g., $2^2 - 0^2 = 4$)

This is not a problem, so nothing needs to be submitted here.

1. Normal, normal, normal [12 points]

For each question below, if you need to use the CDF of a normal, be sure to round the z-score to the hundredths-place and use the table linked on the webpage.

(a) Suppose that X is normally distributed with mean 20 and standard deviation 22. Calculate the probability that $10 < X < 55$.

(b) The weight of a baby at birth is normally distributed with mean 3.2 kilograms and standard deviation 800 grams (approximately). What fraction of babies would you predict weigh more than 4.3 kilograms at birth?

2. Confidence intervals in "real life" [24 points]

In all of the following use the Central Limit Theorem. Use continuity correction if (and only if) you're approximating a discrete random variable.

(a) You are in debt, but luckily you are good at math! You independently owe each of your 30 friends some amount of money by the end of the month, but you aren't sure exactly how much. You do know that you owe each friend an average of $\$135$ with a standard deviation of $\$34$. How much money do you need to be at least 99% sure that you will be able to pay everyone back? Give your answer to the nearest cent. You should treat the amounts of money that you owe as continuous.

(b) After unfortunately failing to pay back your friends, you've decided to turn to other means of earning money: cryptocurrency. After a bit of research, you've narrowed down your focus to 9 different currencies. You're thinking of investing $\$14000$ in every currency. You know you will gross $\$55000$ (for a net return of $\$55000 - \$14000 = \$41000$) with probability p, and with probability $1 - p$ this currency will not be a trading at a price you will want to sell (for a net return of $-\$14000$). Since cryptocurrency trading is based primarily on hype, the 9 cryptocurrencies are mutually independent.

Your local bank has agreed to loan you the money as long as your net return from these investments is positive with at least 95% probability. What is the condition on p that needs to be true in order to secure the loan? You should treat the amount of revenue you'll get from these investments as discrete (since the net return will only increase in multiples of $\$41000$ and $-\$14000$).

3. Exponential in all directions [20 points]

A continuous random variable X has a density function with parameter λ given by:

$$f_X(x) = ce^{-3\lambda|x|} \qquad -\infty < x < \infty,$$

for some constant c. You may use the following facts if needed:
$\lim_{t\to\infty} e^t = \infty$ and $\lim_{t\to-\infty} e^t = 0$.
$\lim_{t\to\infty} e^{-t} = 0$ and $\lim_{t\to-\infty} e^{-t} = \infty$.

(a) If λ is equal to 0 or negative, this is not a valid density function. Explain what property of pdfs is violated when $\lambda \leq 0$. [6 points] We recommend you graph the function on wolframalpha, desmos, or some other graphing calculator for a few values of λ before starting on this question.

For the rest of this problem, assume $\lambda > 0$.

(b) Compute the constant c in terms of λ. [4 points]

(c) Compute the mean and variance of X in terms of λ.
You should use an online calculator like WolframAlpha to evaluate integrals for this question. Specifically, you do **not** need to show (b) and (c) described in Section 0. You **must** show the integral to evaluate and then its final evaluation from the calculator. [5 points]

(d) Compute $Pr(X \geq x)$ in terms of x and λ. (Note that x can be positive or negative or 0. Consider all cases.) [5 points]

4. Joint Continuous Densities [12 points]

Let X and Y be continuous random variables with the following joint distribution.

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{128}xy^3 & 0 \leq x \leq 2, 0 \leq y \leq 4 \\ 0 & \text{otherwise} \end{cases}$$

(a) Find $\mathbb{P}(X > Y)$ [5 points].

(b) What is the marginal PDF of X? [3 points]

(c) What is the marginal PDF of Y? [3 points]

(d) Are X and Y independent? [1 point]

5. Distinct Elements Analysis [5 points]

In this problem you will do some theoretical analysis for the code you will write in Problem 6.

Recall the setup for the problem: YouTube wants to count the number of ***distinct*** views for a video. By "distinct view" we mean the number of different people who have watched a video (so the same person watching a video three times counts only once). You could do this by storing User IDs for every user who has watched the video, but YouTube doesn't want to store all the user ID's, as that would use too much memory. The textbook describes a way to estimate this number.

We modelled the problem as follows: we see a **stream** of 8-byte integers (user ID's), $x_1, x_2, \ldots, x_N$, where x_i is the user ID of the i-th view to a video, but there are only n *distinct* elements ($1 \leq n \leq N$), since some people rewatch the video, even multiple times. We don't know what the number of views N is.

Let $U_1, \ldots, U_m$ be m iid samples from the continuous $\texttt{Unif}(0,1)$ distribution, and let $X = \min\{U_1, \ldots, U_m\}$.

In this problem, you should compute $\text{Var}(X)$. For this problem, you may start with the CDF of X found on page 334 of the textbook, and you may use the fact from the textbook that $\mathbb{E}[X] = \frac{1}{m+1}$. You may find it helpful to do this computation yourself first for practice before moving on to the variance.

You may want to use WolframAlpha to solve any integrals along the way. For this problem, you **do not** need to show the antiderivative but you should **still show the integral to evaluate**.

6. Distinct Elements [15 points]

In this problem we are going to be writing the code to go with the Distinct Elements Analysis.

The scenario for the problem is the following:

YouTube wants to count the number of ***distinct*** views for a video, but doesn't want to store all the user ID's.

We modelled the problem as follows: we see a **stream** of 8-byte integers (user ID's), $x_1, x_2, \ldots, x_N$, where x_i is the user ID of the i^{th} view to a video, but there are only n *distinct* elements ($1 \leq n \leq N$), since some people rewatch the video, even multiple times. We don't know what the number of views N is.

Suppose the universe of user ID's is the set $\mathcal{U}$ (think of this as all 8-byte integers), and we have a single ***uniform*** hash function $h : \mathcal{U} \to [0,1]$. That is, for an integer y, pretend $h(y)$ is a **continuous** $\texttt{Unif}(0,1)$ random variable. That is, $h(y_1), h(y_2), ..., h(y_k)$ for any k **distinct** elements are iid continuous $\texttt{Unif}(0,1)$ random variables, but since the hash function always gives the same output for some given input, if, for example, the i^{th} user ID x_i and the j-th user ID x_j are the same, then $h(x_i) = h(x_j)$ (i.e., they are the "same" $\texttt{Unif}(0,1)$ random variable).

Pseudocode is provided which explains the two key functions that you will implement:

- `UPDATE(x)`: How to update your variable when you see a new stream element.
- `ESTIMATE()`: At any given time, how to estimate the number of distinct elements you've seen so far.

Your task for this problem is to implement the algorithm in python. Starter code is available on the associated Ed lesson.

7. Feedback [1 point]

Answer these questions on the separate Gradescope box for this question.

Please keep track of how much time you spend on this homework and answer the following questions. This can help us calibrate future assignments and future iterations of the course, and can help you identify which areas are most challenging for you.

- How many hours did you spend working on this assignment (excluding any extra credit questions, if applicable)? Report your estimate to the nearest hour.
- Which problem did you spend the most time on?
- Any other feedback for us?