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]". 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.)

Please put each numbered problem on its own page of the pdf (this will make selecting pages easier when you submit), and ensure that your pdfs are oriented correctly (e.g. not upside-down or sideways). The coding problem will also be submitted to gradescope.

Your code will be submitted under "HW 6 [Coding]" as a file called `cse312_pset6_dist_elts.py`.

Due Date: This assignment is due at 11:59 PM Thursday August 5 (Seattle time, i.e. GMT-7).
You will submit the written problems as a PDF to gradescope. Please put each numbered problem on its own page of the pdf (this will make selecting pages easier when you submit), and ensure that your pdfs are oriented correctly (e.g. not upside-down or sideways). The coding problem will also be submitted to gradescope.

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.

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., $2^2 - 0^2 = 4$)

1. The Classic Flea Problem [16 points]

A flea of negligible size is trapped in a large, spherical, inflated beach ball with radius c. At this moment, it is equally likely to be at any point within the ball. Let X be the distance of the flea from the center of the ball. For X, find

(a) the cumulative distribution function F. [5 points]

(b) the probability density function f. [5 points]

(c) the expected value. [4 points]

(d) the variance. [4 points] Reminder: the volume of a sphere of radius c is $\frac{4}{3}\pi c^3$.

2. Exponential Darts [12 points]

You throw a dart at a circular target of radius r. Let X be the distance of your dart's hit from the center of the target. You have improved and your aim is such that $X \sim \text{Exponential}(5/r)$. (Note that it is possible for the dart to completely miss the target.)

(a) As a function of r, determine the value m such that $\mathbb{P}(X < m) = \mathbb{P}(X > m)$. [6 points]

(b) For $r = 14$, give the value of m to 3 decimal places. [2 points]

(c) What is the probability that you miss the target completely? Give your answer to 4 decimal places. [4 points]

3. 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 hundreths-place and use the table linked on the webpage.

(a) Suppose that X is normally distributed with mean 45 and standard deviation 10. Calculate the probability that $20 < X < 50$.

(b) The weight of a baby at birth is normally distributed with mean 3400 grams, with standard deviation 500 grams (approximately). What fraction of babies would you predict weigh more than 4255 grams at birth.

(c) Your friend's rap-battle-bracket was busted shortly after the tournament started, so they asked for another chance to predict the championship battle. You know the abilities of the two performers well, but how they perform on a particular day has some randomness to it. The quality of participant X's rap is normally distributed with mean 5 and variance 2, while participant Y's rap-quality is (independent and) normally distributed with mean 4 and variance 5. What is the probability that person X's rap-quality beats person Y's in the championship?[1] (Hint: you will need to define a new normal random variable).

4. Confidence intervals in "real life" [16 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've decided to scale up your rap-battle-betting ventures. Now, you've enlisted 20 friends who each bet on some amount of rap battles independently and you owe each friend an average of \$127 with a standard deviation of \$35. You currently do not have any money, but luckily, you are in the process of taking out a second mortgage in order to repay your friends. How large of a loan should you take out in order to repay your friends with probability at least 99%. You should treat these amounts of money that you owe as continuous.

(b) After failing to pay back your mortgage, you've decided to turn to other means in order to repay your friends: cryptocurrency. After a bit of research, you've narrowed down your focus to 8 different currencies. You're thinking of investing \$12000 in every currency. You know that each currency will go 'to the moon' with probability p and you will gross \$50000 (for a net profit of \$50000 - \$12000 = \$38000), and with probability $1 - p$ this currency will not be a trading at a price you will want to sell, so you will make a net profit of -\$12000. As cryptocurrency trading is based primarily on hype, the cryptocurrencies are 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 99% probability. What is the condition on p that you need to satisfy in order to secure the loan? You should treat the amount of revenue you'll get from these investments as discrete (since you're only ever adding up multiples of \$38000 and -\$12000).

[1] In this problem, rap quality is measured as a real number. It can be negative or positive. Robbie's raps are always negative quality. The raps in *Hamilton* go as high as 100 in quality

5. Exponential in all directions [20 points]

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

$$f_X(x) = ce^{-2\lambda|x|} \qquad -\infty < x < \infty,$$

for some constant c.

(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. For this part, fully evaluating integrals would require integration by parts. You may skip steps b and c of the standard integral directions. I.e., you may write the integral to evaluate, and skip to the evaluation given by a 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]

6. Distinct Elements [15 points]

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. In lecture, we described a way for them to get a good estimate of this number without storing everything.

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; we can't even store the number n of distinct views (i.e., the number of distinct views).

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).

In class we discussed how to (approximately) solve this distinct elements problem using a single floating point variable (8 bytes), instead of the amount of memory the naive approach from question 5b requires. 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. Distinct Elements Analysis [5 points]

In this problem you will do some theoretical analysis for the code you wrote above.

Recall the setup for the problem: YouTube wants to count the number of ***distinct*** views for a video, but doesn't want to store all the user ID's. In class we described a way for them to get a good estimate of this number without storing everything.

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; we can't even store the number n of distinct views (i.e., the number of distinct views).

Let $U_1, \ldots, U_m$ be m iid samples from the continuous $\texttt{Unif}(0, 1)$ distribution, and let $X = \min\{U_1, \ldots, U_m\}$. We know from lecture (and the textbook) that $\mathbb{E}[X] = \frac{1}{m+1}$. Compute $\text{Var}(X)$. For this problem, you may start with the CDF of X found on page 334 of the textbook.