CSE 312: Foundations of Computing II

Section 4: Practice with Continuous Random Variables

1. Review of Main Concepts

(a) **Cumulative Distribution Function (cdf)**: For any random variable (discrete or continuous) X, the cumulative distribution function is defined as $F_X(x) = \mathbb{P}(X \leq x)$. Notice that this function must be monotonically nondecreasing: if $x < y$ then $F_X(x) \leq F_X(y)$, because $\mathbb{P}(X \leq x) \leq \mathbb{P}(X \leq y)$. Also notice that since probabilities are between 0 and 1, that $0 \leq F_X(x) \leq 1$ for all x, with $\lim_{x \to -\infty} F_X(x) = 0$ and $\lim_{x \to +\infty} F_X(x) = 1$.

(b) **Continuous Random Variable**: A continuous random variable X is one for which its cumulative distribution function $F_X(x) : \mathbb{R} \to \mathbb{R}$ is continuous everywhere. A continuous random variable has an uncountably infinite number of values.

(c) **Probability Density Function (pdf or density)**: Let X be a continuous random variable. Then the probability density function $f_X(x) : \mathbb{R} \to \mathbb{R}$ of X is defined as $f_X(x) = \frac{d}{dx} F_X(x)$. Turning this around, it means that $F_X(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f_X(t)\, dt$. From this, it follows that $\mathbb{P}(a \leq X \leq b) = F_X(b) - F_X(a) = \int_a^b f_X(x) dx$ and that $\int_{-\infty}^{\infty} f_X(x) dx = 1$. From the fact that $F_X(x)$ is monotonically nondecreasing it follows that $f_X(x) \geq 0$ for every real number x.

If X is a continuous random variable, note that in general $f_X(a) \neq \mathbb{P}(X = a)$, since $\mathbb{P}(X = a) = F_X(a) - F_X(a) = 0$ for all a. However, the probability that X is close to a is proportional to $f_X(a)$: for small δ, $\mathbb{P}\left(a - \frac{\delta}{2} < X < a + \frac{\delta}{2}\right) \approx \delta f_X(a)$.

(d) **i.i.d. (independent and identically distributed)**: Random variables $X_1, \ldots, X_n$ are i.i.d. (or iid) if they are independent and have the same probability mass function or probability density function.

(e) **Discrete to Continuous:**

	Discrete	**Continuous**
PMF/PDF	$p_X(x) = \mathbb{P}(X = x)$	$f_X(x) \neq \mathbb{P}(X = x) = 0$
CDF	$F_X(x) = \sum_{t \leq x} p_X(t)$	$F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt$
Normalization	$\sum_x p_X(x) = 1$	$\int_{-\infty}^{\infty} f_X(x)\, dx = 1$
Expectation	$\mathbb{E}[X] = \sum_x x p_X(x)$	$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\, dx$
LOTUS	$\mathbb{E}[g(X)] = \sum_x g(x) p_X(x)$	$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx$

2. Zoo of Continuous Random Variables

(a) **Uniform**: $X \sim \text{Uniform}(a, b)$ iff X has the following probability density function:

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$$

$\mathbb{E}[X] = \frac{a+b}{2}$ and $Var(X) = \frac{(b-a)^2}{12}$. This represents each real number from $[a, b]$ to be equally likely.

(b) **Exponential**: $X \sim \text{Exponential}(\lambda)$ iff X has the following probability density function:

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

$\mathbb{E}[X] = \frac{1}{\lambda}$ and $Var(X) = \frac{1}{\lambda^2}$. $F_X(x) = 1 - e^{-\lambda x}$ for $x \geq 0$. The exponential random variable is the continuous analog of the geometric random variable: it represents the waiting time to the next event,

where $\lambda > 0$ is the average number of events per unit time. Note that the exponential measures how much time passes until the next event (any real number, continuous), whereas the Poisson measures how many events occur in a unit of time (nonnegative integer, discrete). The exponential random variable X is memoryless:

$$\text{for any } s, t \geq 0,\ \mathbb{P}\left(X > s + t \mid X > s\right) = \mathbb{P}(X > t)$$

The geometric random variable also has this property.

(c) **Gamma**: $X \sim \text{Gamma}(r, \lambda)$ iff X has the following probability density function:

$$f_X(x) = \begin{cases} \frac{\lambda^r}{(r-1)!} x^{r-1} e^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

$\mathbb{E}[X] = \frac{r}{\lambda}$ and $Var(X) = \frac{r}{\lambda^2}$. Gamma is the sum of r independent Exp() random variables. Gamma is to Exponential as Negative Binomial to Geometric. It is the waiting time until the r-th event, rather than just the first event. So you can write it as a sum of r independent exponential random variables. It is the waiting time until the rth occurrence of an event in a Poisson Process with parameter λ.

3. Will the battery last?

Suppose that the number of miles that a car can run before its battery wears out is exponentially distributed with expectation 10,000 miles. If the owner wants to take a 5000 mile road trip, what is the probability that she will be able to complete the trip without replacing the battery, given that the car has already been used for 2000 miles?

Solution:

Let N be a r.v. denoting the number of miles until the battery wears out. Then $N \sim \exp(10,000^{-1})$, because N measures the "time" (in this case miles) before an occurrence (the battery wears out) with expectation 10,000. Since this is an exponential distribution, and the expectation of an exponential distribution is $\frac{1}{\lambda}$, $\lambda = \frac{1}{10,000}$. Therefore, via the property of memorylessness of the exponential distribution:

$$\mathbb{P}(N \geq 5000 | N \geq 2000) = \mathbb{P}(N \geq 3000) = 1 - \mathbb{P}(N \leq 3000) = 1 - \left(1 - e^{-\frac{3000}{10000}}\right) \approx 0.741$$

4. Max of uniforms

Let $U_1, U_2, \ldots, U_n$ be mutually independent Uniform random variables on $(0, 1)$. Find the CDF and PDF for the random variable $Z = \max(U_1, \ldots, U_n)$.

Solution:

The key idea for solving this question is realizing that the max of n numbers $\max(a_1, ..., a_n)$ is less than some constant c, if and only if each individual number is less than that constant c (i.e. $a_i < c$ for all i). Using this idea, we get

$$\begin{aligned} F_Z(x) = \mathbb{P}(Z \leq x) &= \mathbb{P}(\max(U_1, ..., U_n) \leq x) \\ &= \mathbb{P}(U_1 \leq x, ..., U_n \leq x) \\ &= \mathbb{P}(U_1 \leq x) \cdot ... \cdot \mathbb{P}(U_n \leq x) && \text{[independence]} \\ &= F_{U_1}(x) \cdot ... \cdot F_{U_n}(x) \\ &= F_U(x)^n && \text{[where } U \sim \text{Unif}(0,1)\text{]} \end{aligned}$$

So the CDF of Z is

$$F_Z(x) = \begin{cases} 0 & x < 0 \\ x^n & 0 \leq x \leq 1 \\ 1 & x > 1 \end{cases}$$

To find the PDF, we take the derivative of each part of the CDF, which gives us the following

$$f_Z(x) = \begin{cases} n\, x^{n-1} & 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

5. New PDF?

Alex came up with a function that he thinks could represent a probability density function. He defined the potential pdf for X as $f(x) = \frac{1}{1+x^2}$ defined on $[0, \infty)$. Is this a valid pdf? If not, find a constant c such that the pdf $f_X(x) = \frac{c}{1+x^2}$ is valid. Then find $\mathbb{E}[X]$. (Hints: $\frac{d}{dx}(\tan^{-1} x) = \frac{1}{1+x^2}$, $\tan \frac{\pi}{2} = \infty$, and $\tan 0 = 0$.)

Solution:

The area under the PDF is 1. So,

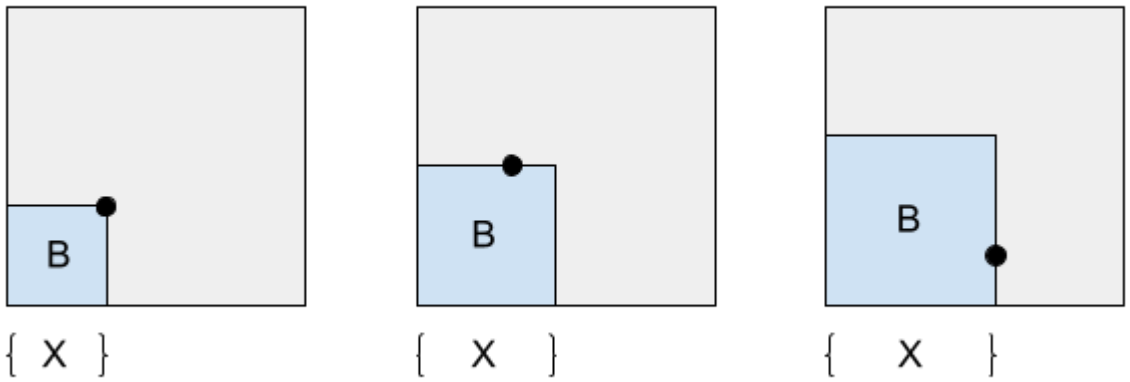
$$\int_0^\infty \frac{c}{1+x^2} dx = c \tan^{-1} x \,|_0^\infty = c\left(\frac{\pi}{2} - 0\right) = 1$$

Solving for c gives us $c = 2/\pi$. Using our value we found for c, and the definition of expectation we can compute $E[X]$ as follows:

$$\mathbb{E}[X] = \int_0^\infty \frac{cx}{1+x^2} dx = \frac{2}{\pi} \int_0^\infty \frac{x}{1+x^2} dx = \frac{1}{\pi} \ln(1+x^2) \,|_0^\infty = \infty$$

6. A square dartboard?

You throw a dart at an $s \times s$ square dartboard. The goal of this game is to get the dart to land as close to the lower left corner of the dartboard as possible. However, your aim is such that the dart is equally likely to land at any point on the dartboard. Let random variable X be the length of the side of the smallest *square* B in the lower left corner of the dartboard that contains the point where the dart lands. That is, the lower left corner of B must be the same point as the lower left corner of the dartboard, and the dart lands somewhere along the upper or right edge of B. See the image below for three examples of how X can take on a value.



For X, find the CDF, PDF, $\mathbb{E}[X]$, and $Var(X)$.

Solution:

Since $F_X(x)$ is the probability that the dart lands inside the square of side length x, that probability is the area of a square of length x divided by the area of the square of length radius s (i.e., x^2/r^2). Thus, our CDF looks like

$$F_X(x) = \begin{cases} 0, & \text{if } x < 0 \\ x^2/s^2, & \text{if } 0 \le x \le s \\ 1, & \text{if } x > s \end{cases}$$

To find the PDF, we just need to take the derivative of the CDF, which gives us the following:

$$f_X(x) = \frac{d}{dx}F_X(x) = \begin{cases} 2x/s^2, & \text{if } 0 \le x \le s \\ 0, & \text{otherwise} \end{cases}$$

Using the definition of expectation and variance we can compute $\mathbb{E}[X]$ and $Var(X)$ in the following manner:

$$\mathbb{E}[X] = \int_0^s x f_X(x)dx = \int_0^s \frac{2x^2}{s^2}dx = \frac{2}{s^2}\int_0^s x^2 dx = \frac{2}{3s^2}\left[x^3\right]_0^s = \frac{2}{3}s$$

$$\mathbb{E}\left[X^2\right] = \int_0^s x^2 f_X(x)dx = \int_0^s \frac{2x^3}{s^2}dx = \frac{2}{s^2}\int_0^s x^3 dx = \frac{1}{2s^2}\left[x^4\right]_0^s = \frac{1}{2}s^2$$

$$\text{Var}(X) = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2 = \frac{1}{2}s^2 - \left(\frac{2}{3}s\right)^2 = \frac{1}{18}s^2$$

7. Gender composition of classes

[Credit to Chris Piech, Stanford CS109] A massive online class has sections with 10 students each. Each student in our population has a 50% chance of identifying as female, 47% chance of identifying as male and 3% chance of identifying as non-binary. Even though students are assigned randomly to sections, a few sections end up having a very uneven distribution just by chance. You should assume that the population of students is so large that the percentages of students who identify as male / female / non-binary are unchanged, even if you select students without replacement.

(a) Define a random variable for the number of people in a section who identify as female.

(b) What is the expectation and standard deviation of number of students who identify as female in a single section?

(c) Write an expression for the exact probability that a section is skewed. We defined skewed to be that the section has 0, 1, 9 or 10 people who identify as female.

(d) The course has 1,200 sections. Approximate the probability that at 5 or more sections will be skewed.

Solution:

(a) Let X denote the number of people in a section who identify as female. $X \sim \text{Bin}(n = 10, p = 0.5)$.

(b) $\mathbb{E}[X] = n \cdot p = 10 \cdot 0.5 = 5$
$\text{Std}(X) = \sqrt{Var(X)} = \sqrt{n \cdot p \cdot (1-p)} = \sqrt{10 \cdot 0.5 \cdot 0.5} \approx 1.6$

(c) Recall that $p = 0.5$
$P(\text{skewed}) = P(X = 0) + P(X = 1) + P(X = 9) + P(X = 10) =$
$\binom{10}{0}(1-p)^{10} + \binom{10}{1}p(1-p)^9 + \binom{10}{9}p^9(1-p) + \binom{10}{10}p^{10} \approx 0.021$

(d) The exact probability of number of skewed sections is $S \sim \text{Bin}(n = 1200, p = 0.021)$. This will require excessive calculations to reason about. Instead, we can approximate the number of skewed sections using a Poisson approximation. Let Y be the Poisson approximation of S.

$Y \sim \text{Poi}(\lambda = 25.2)$ since $n \cdot p = 1200 \cdot 0.021 = 25.2$

$$\begin{aligned} P(Y > 5) &= 1 - P(Y < 5) \\ &= 1 - (P(Y = 0) + P(Y = 1) + P(Y = 2) + P(Y = 3) + P(Y = 4)) \\ &> 0.9999 \end{aligned}$$

8. Website visits

You have a website where only one visitor can be on the site at a time, but there is an infinite queue of visitors, so that immediately after a visitor leaves, a new visitor will come onto the website. On average, visitors leave your website after 5 minutes. Assume that the length of stay is exponentially distributed. What is the probability that a user stays more than 10 minutes, if we calculate this probability:

(a) Using the random variable X defined as the length of stay of the user?

(b) Using the random variable Y, defined as the number of users who leave your website over a 10-minute interval?

Solution:

(a) $X \sim \text{Exp}(\lambda = 0.2)$

$$P(X > 10) = 1 - F_X(10) = 1 - (1 - e^{-10\lambda}) = e^{-2} \approx 0.1353$$

(b) $Y \sim \text{Poi}(\lambda = 2)$

$$P(Y = 0) = \frac{2^0 e^{-2}}{0!} = e^{-2} \approx 0.1353$$