Section 01: Probability Review

1. PDF, CDF and Expectation

The **Probability Density Function** (PDF), or probability mass function, $f_X : \mathbb{R} \to \mathbb{R}^{\geq 0}$ of a random variable X is defined as $\mathbb{P}(X = x)$. The **Cumulative Density Function** (CDF) $F_X : \mathbb{R} \to [0, 1]$ of that same random variable is defined as $\mathbb{P}(X \leq x)$.

Note that the CDF can be computed from the PDF, and vice versa; e.g. $F_X = \int_{-\infty}^{x} f(x)dx$.

We can use these functions to directly compute the expectation of random variables, since the expectation is defined in terms of the PDF: $\mathbb{E}(X) = \int_{-\infty}^{\infty} x \cdot f_X(x)dx$.

These functions can also be used to compute the distribution of any one-to-one transformation $g(\cdot)$ of the random variable: $\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x) \cdot f_X(x)dx$.

Note: this section focuses on the continuous case, but equivalent formulations hold in the discrete case by replacing integration with summation.

(a) You've just started a new exercise regimen. You start on the 2nd floor of CSE1, and then make a random choice:

- With probability p_1 you run up 2 flights of stairs.
- With probability p_2 you run up 1 flight of stairs.
- With probability p_3 you walk down 1 flight of stairs.

Where $p_1 + p_2 + p_3 = 1$.

You will do two iterations of your exercise scheme (with each draw being independent). Let X be the floor you're on at the end of your exercise routine. Recall you start on floor 2.

(i) Let Y be the expected difference between your ending floor and your starting floor in one iteration. What is $\mathbb{E}[Y]$ (in terms of p_1, p_2, p_3)?

(ii) What is $\mathbb{E}[X]$ (use your answer from the previous part)

(iii) You change your scheme: instead of doing two independent iterations, you decide the second iteration of your regimen will just use the same random choice as your first (in particular they are no longer independent!). Does $\mathbb{E}[X]$ change? (Optional)

Fact 1. *Let* $X_{(j)}$ *denote the* j*th order statistic in a sample of i.i.d. random variables; that is, the* j*th element when the items are sorted in increasing order* $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$*.*

The PDF of $X_{(j)}$ *is given by:*

$$f_{X_{(j)}}(x) = \frac{n!}{(n-j)!(j-1)!}[F(x)]^{j-1}[1-F(x)]^{n-j}f(x). \tag{1}$$

(b) When a sample of $2n+1$ i.i.d. random variables is observed, the $(n+1)^{\text{st}}$ smallest is called the sample median. If a sample of size 3 from a uniform distribution over $[0,1]$ is observed, find the probability that the sample median is between $\frac{1}{4}$ and $\frac{3}{4}$. *Hint: use Fact 1.*

2. Linearity and Independence

Suppose we have two random variables X and Y, such that $\mathbb{E}[X] = \mathbb{E}[Y] = 2$. For each of the following quantities either:

- State the value of the quantity if we have enough information to find it, or
- Give examples of two different values the quantity could take if we do not.

(a) $\mathbb{E}[X + Y]$

(b) $\mathbb{E}[XY]$

(c) $\mathbb{E}[X^2]$

(d) $\mathbb{E}[X]^2$

Suppose we additionally know that X and Y are independent. Do any of the answers change?

3. Variance and Concentration

Sewoong wants to see if the students in the course like probability theory. You (because you're so friendly) know that 200 out of the 250 students in the course say they like probability theory, but Sewoong doesn't believe you. They decide to use the following process to estimate the number of people who like probability theory:

- Choose a student uniformly at random (and independent from any previous choices).
- Record $X_i = \begin{cases} 1 & \text{if the student likes probability} \\ 0 & \text{otherwise} \end{cases}$

They will choose 30 such students this way, and they define $X = \frac{\sum_{i=1}^{30} X_i}{30}$, the average of the X_i.

(a) What is $\mathbb{E}[X_1]$?

(b) What is Var(X_1)? Hint: $p(1-p)$ is the variance of a Bernoulli random variable with probability of success p.

(c) What is $\mathbb{E}[X]$?

(d) What is Var(X)?

Theorem 1 (Chebyshev's Inequality). *If X is a random variable with finite mean μ and finite variance σ^2, then for any real number $k > 0$:*

$$\mathbb{P}\left[|X - \mu| \geq k\sigma\right] \leq \frac{1}{k^2}$$

(e) Sewoong is worried that less than half the course likes probability theory. They will stop being worried if $X \geq 0.5$. Use Chebyshev's inequality to give a lower bound on the probability that they stop worrying.