EE 520: Random Processes **Fall 2021**

Lecture 3
Discrete Random Variables and Expectation

Instructor Name: John Lipor

Recommended Reading: Pishro-Nik: 3.1 - 3.2; Gubner: 2.1 - 2.4

1 Random Variables

We are often interested in functions of events our outcomes, rather than individual events/outcomes themselves. The function that maps outcomes ω in the sample space Ω is called a *random variable*.

Example 1. Toss a fair coin twice, yielding the sample space $\Omega = \{HH, HT, TH, TT\}$. For some outcome $\omega \in \Omega$, let $X(\omega)$ be the number of heads, which gives

$$X(HH) = 2 \quad X(HT) = X(TH) = 1 \quad X(TT) = 0.$$

Definition 1. A **random variable** (RV) is a function $X : \Omega \to \mathbb{R}$ such that

$$\{\omega \in \Omega : X(w) \leq x\} \in \mathcal{F}$$

for every $x \in \mathbb{R}$, i.e., X is $\mathcal{F}$-measurable.

Note that we often omit the dependence on ω and simply write X. The measurability condition above is a technical condition that ensures the cumulative distribution function (CDF, more on this later) exists. Recall that a probability measure P measures the size of sets from our sample space Ω. If we want to talk about the probability that a random variable lies below some value x, then we need to make sure we can measure the corresponding set of outcomes.

Example 2. In the first example, what is the probability we get at least one head? In terms of the RV X, we want to know

$$\begin{aligned} P(X \geq 1) &= P(\{\omega \in \Omega : X(\omega) \geq 1\}) \\ &= P(\{HT, TH, HH\}) = \frac{3}{4}. \end{aligned}$$

It can be useful to discuss how likely different values of X are, i.e., to find $P(X = x)$. However, a more useful/general function is called the *cumulative distribution function*.

Definition 2. For a RV $X : \Omega \to \mathbb{R}$, the **cumulative distribution function** (CDF), also referred to as simply the **distribution**, is a function $F : \mathbb{R} \to [0, 1]$ defined by

$$F(x) = P(X \leq x) = P(\mathcal{A}(x)),$$

where $\mathcal{A}(x) = \{\omega \in \Omega : X(\omega) \leq x\}$.

Note that to make statements about $P(\mathcal{A}(x))$, the set $\mathcal{A}(x)$ must be in the σ-algebra $\mathcal{F}$, which is why we need the measurability condition in the definition of a RV. We will now look at some important examples of RVs.

2 Discrete Random Variables

You likely have an intuitive understanding of what a discrete RV is. Below is a formal definition.

Definition 3. The RV $X : \Omega \to \mathbb{R}$ is called **discrete** if it takes values in some countable subset $\{x_1, x_2, \dots\} \subset \mathbb{R}$ only.

For discrete RVs, we can measure the probability of taking a specific value.

Definition 4. The **probability mass function** (PMF) of a discrete RV is

$$p_X(x) = P(X(\omega) = x),$$

and it holds that

$$\sum_i p_X(x_i) = 1,$$

where the summation is over all possible values of X.

2.1 Common random variables (a.k.a., the "big 5")

- **uniform**: "Equally likely" or "random" events are drawn from the uniform distribution.

$$P(X = k) = \frac{1}{n}, \quad k = 1, \dots, n$$
$$\Leftrightarrow p_X(k) = \begin{cases} \frac{1}{n}, & k = 1, \dots, n \\ 0, & \text{otherwise.} \end{cases}$$

- **Bernoulli**: Event happens with probability p. Taking $\Omega = \{0, 1\}$ and $\mathcal{F} = 2^\Omega$ (set of all subsets of Ω), this has the PMF

$$p_X(1) = P(\{1\}) = p$$
$$p_X(0) = p(\{0\}) = 1 - p.$$

- **binomial**: k successes in n trials, where each success occurs with probability p.

$$p_X(k) = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, \dots, n.$$

- **geometric**: Number of realizations/trials before first success.

$$p_X(k) = P(X = k) = p(1-p)^{k-1}.$$

  Note: There is some discrepancy as to how the geometric distribution is defined, so be careful when using Wikipedia or other sources.

- **Poisson**: Models many physical phenomena, especially arrival processes.

$$p_X(k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \dots.$$

3 Multiple Random Variables

Multiple RVs give us a shorthand for talking about multiple functions of outcomes simultaneously. In this case, we talk about probabilities of the form

$$P(X \in B, Y \in C) = P\left(\{\omega \in \Omega : X(\omega) \in B\} \cap \{\omega \in \Omega : Y(\omega) \in C\}\right).$$

Definition 5. Two RVs are **independent** if the events

$$\{\omega \in \Omega : X(\omega) \in B\} \text{ and } \{\omega \in \Omega : Y(\omega) \in C\}$$

are independent for all $B, C \in \mathcal{F}$, i.e., if

$$P(X \in B, Y \in C) = P(X \in B)P(Y \in C).$$

Definition 6. Multiple RVs are called **identically distributed** if $P(X_j \in B)$ does not depend on j. If a set of RVs are independent and identically distributed, we say they are i.i.d.

When dealing with multiple RVs, we can think about their joint PMF and use it to derive the individual (marginal) PMFs.

Definition 7. For RVs X, Y, the **joint PMF** is

$$p_{XY}(x_i, y_j) = P(X = x_i, Y = y_j).$$

Definition 8. For two RVs X, Y with joint PMF p_{XY}, the **marginal PMF** of X is defined as

$$p_X(x_i) = \sum_j p_{XY}(x_i, y_j)$$

and similar for Y.

Note that by the definition of the joint PMF, two RVs are independent if and only if the joint distribution factors, i.e., if

$$p_{XY}(x_i, y_j) = p_X(x_i)p_Y(y_j).$$

4 Expectation

Expectation generalizes the notion of "average" you learned in elementary school. This was the *sample mean*

$$m = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

which we will see corresponds to the expectation with respect to the uniform distribution.

Definition 9. The **expectation** or **expected value** of a discrete RV is

$$\mathbb{E}[X] = \sum_i x_i p_X(x_i).$$

Example 3. Let $X \sim \text{Ber}(p)$ (X is a Bernoulli RV with parameter p). Then

$$\mathbb{E}[X] = \sum_{i=1}^{2} x_i p_X(x_i) = 1 \times p + 0 \times (1-p) = p.$$

We can compute expectations of functions using the law of the unconscious statistician (LOTUS). For a function $g : \mathbb{R} \to \mathbb{R}$, we have

$$\mathbb{E}[g(X)] = \sum_i g(x_i) p_X(x_i).$$

Example 4.

$$\mathbb{E}[aX] = \sum_i a x_i = a \sum_i x_i p_X(x_i) = a\mathbb{E}[X].$$

Proposition 1. Expectation is **linear**, i.e.,

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y].$$

Proof. Use LOTUS and try it yourself :) □

The expectation of a RV is called the *first moment* or first-order statistic and is loosely the most important thing about a RV. The next most important statistic uses the *second moment* $\mathbb{E}[X^2]$.

Definition 10. The **variance** of a RV X is

$$\mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2.$$

Variance provides an idea of deviation from the mean and is the square of the standard deviation.

5 Indicator Functions

The indicator function is a very useful tool in probability and central to the analysis of machine learning algorithms. Understanding how to correctly use indicators requires working several examples, but they are an important tool worth developing.

Definition 11. The **indicator function** is defined as

$$\mathbb{1}_A(x) = \begin{cases} 1, & x \in A \\ 0, & x \notin A. \end{cases}$$

Intuitively, an indicator simply indicates whether the variable x lies in the set A. Indicators become useful in probability for the following reason. If we take Ω as the domain of the indicator, we get $\mathbb{1} : \Omega \to \{0, 1\}$, meaning the indicator is a valid RV. Its expectation is

$$\mathbb{E}[\mathbb{1}_A] = 1 \times P\left(\{\omega \in \Omega : \omega \in A\}\right) + 0 \times P\left(\{\omega \in \Omega : \omega \notin A\}\right) = P(A).$$

In other words, the indicator gives us a way to switch between thinking about probabilities and expectations. More generally, if X is a RV, then we can take $\mathbb{1}_A(X)$ to see that

$$\mathbb{E}[\mathbb{1}_A(X)] = P\left(\mathbb{1}_A(X) = 1\right) = P(X \in A).$$

Example 5. Suppose N people throw their hats into the center of a room. The hats are mixed up, and everyone selects one at random. What is the expected number of people who will find their own hat?

Let X be the number of matches, and define the indicator RV

$$X_i = \begin{cases} 1, & \text{if person } i \text{ selects their own hat} \\ 0, & \text{otherwise,} \end{cases}$$

which implies

$$X = \sum_{i=1}^{N} X_i.$$

We want $\mathbb{E}[X]$, and by linearity of expectation we have

$$\mathbb{E}[X] = \sum_{i=1}^{N} \mathbb{E}[X_i].$$

Therefore, our main task is to compute $\mathbb{E}[X_i] = P(X_i) = 1$. Since each person is equally likely to select their own hat, so

$$\mathbb{E}[X_i] = P(X_i = 1) = \frac{1}{N}.$$

We can then easily compute

$$\mathbb{E}[X] = \sum_{i=1}^{N} \mathbb{E}[X_i] = N \times \frac{1}{N} = 1.$$

General strategy for using indicators: When looking for some $\mathbb{E}[X]$ that is hard to compute, attempt the following sequence of steps.

1. Define the indicator RV X_i such that $X = \sum_{i=1}^{N} X_i$
2. Compute $P(X_i = 1)$
3. $\mathbb{E}[X] = \sum_{i=1}^{N} \mathbb{E}[X_i] = \sum_{i=1}^{N} P(X_i = 1)$.