1. Random Variables

In this section we review some properties of random variables from an introductory probability class.

1.1. Probability Spaces.

To formally define a random variable, we begin by introducing the notion of a **probability space**. Probability spaces consist of three parts:

(1) A sample space Ω that contains all possible outcomes ω of some experiment or random trial.
(2) An event space $\mathscr{F}$ that contains all of the events that we can assign probabilities to. Namely, $\mathscr{F}$ is the collection of subsets $A \subseteq \Omega$ that we can assign probabilities to.
(3) A probability measure $\mathbb{P}$ that assigns a probability $\mathbb{P}(A)$ to each event $A \in \mathscr{F}$.

We refer to the triplet $(\Omega, \mathscr{F}, \mathbb{P})$ as a probability space.

Definition 1.1. *A function $\mathbb{P} : \mathscr{F} \to [0,1]$ is a **probability measure** if it satisfies:*

(1) For each $A \in \mathscr{F}$, $\mathbb{P}(A) \in [0,1]$ (the probability of each event is some number between 0 and 1).
(2) $\mathbb{P}(\Omega) = 1$ (the probability that some *outcome takes place is 1).*
(3) If $\{A_n\}_{n=1}^{\infty} = \{A_1, A_2, \ldots\}$ is a collection of pairwise disjoint events (namely, $A_i \cap A_j = \emptyset$ whenever $i \neq j$), then

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(A_n).$$

Some important properties of probability measures that follow from Definition 1.1 are:

(1) For each $A \in \mathscr{F}$, $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.
(2) For each $A, B \in \mathscr{F}$, $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.[1]
(3) If $A, B \in \mathscr{F}$ are such that $A \subseteq B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$.
(4) $\mathbb{P}(\emptyset) = 0$.

It is a helpful exercise to remind yourself why each of these properties follows from the definition above.

Example 1.2. *You have a coin that is equally likely to land on heads and tails.*

If you flip the coin once, then the probability space $(\Omega_1, \mathscr{F}_1, \mathbb{P}_1)$ is given by $\Omega_1 = \{H, T\}, \mathscr{F}_1 = \{\{H, T\}, \{H\}, \{T\}, \emptyset\}$, and the probability measure $\mathbb{P}$ is defined by $\mathbb{P}(\{H\}) = \mathbb{P}(\{T\}) = 1/2$.

If you flip the coin twice, then the probability space $(\Omega_2, \mathscr{F}_2, \mathbb{P}_2)$ is given by

$$\Omega_2 = \{(H,T), (T,H), (H,H), (T,T)\},$$

and $\mathscr{F}_2$ contains all subsets of Ω_2. For instance, $\mathscr{F}_2$ contains the event that you "flip heads then tails or flip tails then tails". We denote this event by $\{(H,T), (T,T)\}$. Here the probability measure on Ω_2 is given by

$$\mathbb{P}(\{(H,T)\}) = P(\{(T,H)\}) = P(\{(H,H)\}) = P(\{(T,T)\}) = 1/4.$$

Recall that two events A and B are **independent** if and only if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$

We often are not interested in the outcomes themselves, but rather in some numerical value derived from the outcomes. For example, when we flip two coins we might be interested in the *number* of coins that land on heads, rather than the order in which the coins land. If we let X denote the number of coins that land on heads, then X assigns a numerical value to each outcome in the sample space. Namely, it is a *function* defined on the sample space. By that we mean that for each $\omega \in \Omega_2$, there is a corresponding value $X(\omega)$ that tells us how many coins landed on heads. Here we can precisely write down how the function $X : \Omega_2 \to \{0,1,2\}$ is defined:

- $X(\{(H,H)\}) = 2$

[1] Recall that $A \cup B \doteq$ "A or B" and $A \cap B \doteq$ "A and B". For a quick review of set notation, see `https://www.purplemath.com/modules/setnotn.htm`.

- $X(\{(H,T))\}) = 1$
- $X(\{(T,H)\}) = 1$
- $X(\{(T,T)\}) = 0$

The function X defined above is an example of a **random variable**. We think of random variables as functions that input an outcome (i.e., some scenario) and output a number. We refer to the possible numbers that the random variable X can take on as the **state space** of X.

Some important notation regarding random variables:

(1) Since we think of X as a function, it has a *domain* and *range*. Its domain is the sample space Ω, and its range is the state space, which we typically denote by $\mathscr{S}_X$. Note that
$$\mathscr{S}_X = X(\Omega) \doteq \{x \in \mathbb{R} : \text{ there is some } \omega \in \Omega \text{ such that } X(\omega) = x\}$$
(2) For a subset $A \subseteq \mathscr{S}_X$, we write
$$\{X \in A\} \doteq \{\omega \in \Omega : X(\omega) \in A\}.$$
Note that $\{X \in A\} \subseteq \Omega$.
(3) For a subset $A \subseteq \mathscr{S}_X$, we write
$$\mathbb{P}(X \in A) \doteq \mathbb{P}(\{X \in A\}) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in A\}).$$
(4) Random variables are generally denote by capital letters such as X, Y, Z, etc.

Example 1.3. *You roll a fair four-sided die twice. Let X denote the maximum of the two rolls. Write down the probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and the state space $\mathscr{S}_X$, and specify X explicitly as a function from the sample space Ω to the state space $\mathscr{S}_X$.*

You may not have seen random variables formulated this way before. Typically it will suffice to think of a random variable as exactly that, namely a random numerical quantity. However, it will occasionally be helpful for us to think of random variables as functions defined on some sample space.

1.2. **Discrete Random Variables.** A **discrete random variable** is a random variable that can take on only an enumerable number of values. Some common state spaces of discrete random variables are:

(1) $\mathscr{S}_X = \{1,2,3,4,5,6\}$
(2) $\mathscr{S}_X = \{0,1\}$
(3) $\mathscr{S}_X = \mathbb{N} \doteq \{1,2,3,\ldots\}$
(4) $\mathscr{S}_X = \mathbb{N}_0 \doteq \{0,1,2,\ldots\}$

Discrete random variables are described in terms of a **probability mass function** (p.m.f.), which is a function $p_X : \mathscr{S}_X \to [0,1]$ defined as
$$p_X(x) = \mathbb{P}(X = x).$$
Note that if p_X is the p.m.f. of a random variable X, then
$$\sum_{x \in \mathscr{S}_X} p_X(x) = \sum_{x \in \mathscr{S}_X} \mathbb{P}(X = x) = 1,$$
and for each $A \subseteq \mathscr{S}_X$,
$$\mathbb{P}(X \in A) = \sum_{x \in A} \mathbb{P}(X = x) = \sum_{x \in A} p_X(x).$$
Some common discrete random variables are Bernoulli, binomial, Poisson, geometric, etc.

Example 1.4. *Consider the random variable X from Example* 1.3. *The p.m.f. of X is given by*
$$p_X(x) = \begin{cases} \frac{1}{16} & x = 1 \\ \frac{3}{16} & x = 2 \\ \frac{5}{16} & x = 3 \\ \frac{7}{16} & x = 4. \end{cases}$$

To calculate the probability that X is between 1 and 3, we evaluate

$$\mathbb{P}(1 \le X \le 3) = \sum_{x=1}^{3} p_X(x) = \frac{1}{16} + \frac{3}{16} + \frac{5}{16} = \frac{9}{16}.$$

1.3. **Continuous Random Variables.** A **continuos random variable** is a random variable that can take on an uncountable number of values. Some common state spaces of continuous random variables are:

(1) $\mathscr{S}_X = \mathbb{R} = (-\infty, \infty)$
(2) $\mathscr{S}_X = \mathbb{R}_+ = [0, \infty)$
(3) $\mathscr{S}_X = [0, 1]$

Continuous random variables are described in terms of a **probability density function** (p.d.f.), which is a function $f_X : \mathbb{R} \to [0, \infty)$ such that

$$\int_{-\infty}^{\infty} f_X(x)\,dx = 1.$$

The idea is that if f_X is relatively large in some region, then it is more likely that X will take on a value in that region. In particular, for a continuous random variable X, to compute probabilities we look at

$$\mathbb{P}(X \in A) = \int_A f_X(x)\,dx.$$

Note that for each $c \in \mathbb{R}$,

$$\mathbb{P}(X = c) = \int_c^c f_X(x)\,dx = 0,$$

so $f_X(c) \neq \mathbb{P}(X = c)$. Generally we will compute probabilities of the form

$$\mathbb{P}(X \in (a,b)) = \mathbb{P}(a < X < b) = \mathbb{P}(X \in [a,b]) = \mathbb{P}(a \le X \le b) = \int_a^b f_X(x)\,dx.$$

Some common example of continuous random variables are uniform, exponential, normal, etc.

Example 1.5. *For $\lambda > 0$, we say that $X \sim Exp(\lambda)$ if*

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \ge 0 \\ 0 & x < 0. \end{cases}$$

Note that $\mathscr{S}_X = \mathbb{R}_+$. To calculate $\mathbb{P}(X \in (a,b))$, where $0 \le a \le b$, we use to change of variables $u \doteq \lambda x$, $du \doteq \lambda\,dx$, to see

$$\begin{aligned} \mathbb{P}(X \in (a,b)) &= \int_a^b f_X(x)\,dx \\ &= \int_a^b \lambda e^{-\lambda x}\,dx \\ &= \int_{\lambda a}^{\lambda b} e^{-u}\,du \\ &= -e^{-u}\Big|_{\lambda a}^{\lambda b} \\ &= e^{-\lambda a} - e^{-\lambda b} \end{aligned}$$

1.4. **Cumulative Distribution Functions.** The **probability distribution** of a random variable describes how likely the different values of a random variable are. Two discrete random variables have the same p.m.f. if and only if they have the same probability distribution. However, the same is not true for continuous random variables. For

example, consider $X \sim \text{Exp}(1)$, and consider the random variable Y whose p.d.f. is given by

$$f_Y(y) = \begin{cases} e^{-y} & y \in \mathbb{R}_+ \setminus \{1\} \\ 10 & y = 1 \\ 0 & y < 0. \end{cases}$$

Then $\mathscr{S}_X = \mathscr{S}_Y = \mathbb{R}_+$, and for any $0 \le a < b$ we have that

$$\mathbb{P}(a \le X \le b) = \mathbb{P}(a \le Y \le b) = e^{-a} - e^{-b}.$$

This should tell us that X and Y have the same probability distribution, even though their p.m.f.'s are different. This motivates the definition of another function, defined in terms of a random variable, that completely describes its probability distribution.

Definition 1.6. *For a random variable X, the function $F_X : \mathbb{R} \to [0,1]$ defined as*

$$F_x(x) \doteq \mathbb{P}(X \le x)$$

*is the **cumulative distribution function** (c.d.f.) of X.*

Below we list several important properties of c.d.f.'s.

Remark 1.7. *(1) If X is a discrete random variable with p.m.f. p_X, then*

$$F_X(x) \doteq \sum_{y \le x} p_X(y). \tag{1}$$

(2) If X is a discrete random variable with p.d.f. f_X, then

$$F_X(x) \doteq \int_{-\infty}^{x} f_X(y)\,dy. \tag{2}$$

(3) For any random variable X, the function F_X is non-decreasing and right-continuous, namely the following properties hold:
- *(a) If $x \le y$, then $F_X(x) \le F_Y(y)$.*
- *(b) For each $x \in \mathbb{R}$, $\lim_{z \to x^+} F_X(z) = F_X(x)$.*

(4) The following limits hold:
- *(a) As $x \to -\infty$, $F_X(x) \to 0$.*
- *(b) As $x \to \infty$, $F_X(x) \to 1$.*

(5) The c.d.f. of a random variable characterizes the distribution/probability law of a random variable uniquely. Namely, if X and Y are random variables with c.d.f.'s F_X and F_Y, respectively, then we say that X and Y have the same probability distribution if $F_X(z) = F_Y(z)$ for all $z \in \mathbb{R}$. In that case, we write $X \overset{d}{=} Y$ or $X \overset{\mathscr{L}}{=} Y$.

Example 1.8. *Let $X \sim Exp(\lambda)$ be as in Example* 1.5. *Note that the state space of X is $\mathscr{S}_X = \mathbb{R}_+$, so if $x < 0$, then*

$$F_X(x) = \mathbb{P}(X \le x) = 0,$$

and if $x \ge 0$, then

$$F_X(x) \doteq \mathbb{P}(X \le x) = \mathbb{P}(0 \le X \le x) = \int_0^x \lambda e^{-\lambda y}\,dy = \int_0^{\lambda x} e^{-u}\,du = -e^{-u}\Big|_0^{\lambda x} = 1 - e^{-\lambda x}.$$

Above we used the change of variables of $u \doteq \lambda y$. Therefore c.d.f. of X is the function $F_X : \mathbb{R} \to [0,1]$ given by

$$F_X(x) = \begin{cases} 1 - e^{-\lambda x} & x \ge 0 \\ 0 & x < 0. \end{cases}$$

The next example is slightly more involved. It will be important when you study continuous time Markov chains. Recall that we say that two random variables X and Y are **independent** if for all events A and B we have

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B).$$

Example 1.9. *Helen and Teddy are waiting to be fed after I wake up in the morning. However, they don't always get fed at the same time. Their waiting times X and Y (in hours) are modeled as independent, Exp(λ) random variables.*

Let M denote the amount of time that it takes for the first *cat to be fed, namely $M \doteq \min\{X, Y\}$. What is the distribution/probability law of M.*

In order to solve this problem, note that we can write

$$\{M > x\} = \{\min\{X, Y\} > x\} = \{X > x, Y > x\}.$$

Since X and Y are independent, we know from Example 1.8 *that*

$$\mathbb{P}(X > x, Y > x) = \mathbb{P}(X > x)\mathbb{P}(Y > x) = (1 - F_X(x))(1 - F_Y(x)) = e^{-\lambda x}e^{-\lambda x} = e^{-2\lambda x}$$

Therefore

$$\begin{aligned} F_M(x) &= \mathbb{P}(M \le x) \\ &= \mathbb{P}(\min\{X, Y\} \le x) \\ &= 1 - \mathbb{P}(\min\{X, Y\} > x) \\ &= 1 - \mathbb{P}(X > x, Y > x) \\ &= 1 - e^{-2\lambda x}. \end{aligned}$$

Since the c.d.f. uniquely characterizes a random variable's probability distribution, it follows that $M \sim Exp(2\lambda)$.

If instead of two cats, I had n cats, and their feeding times in the morning, denoted by $X_1, X_2, \ldots, X_n$, were all independent Exp(λ) random variables, then what would the distribution of the first feeding time $M \doteq \min\{X_1, X_2, \ldots, X_n\}$ be?

The main takeaway of this section is that the c.d.f. of a random variable is a function that fully describes the random variable's probability distribution.

1.5. **Expected Value.** The expected value of a random variable describes its average value or mean.

Definition 1.10. *A discrete random variable X with p.m.f. p_X and state space $\mathscr{S}_X$ is **integrable** if*

$$\sum_{x \in \mathscr{S}_X} |x| p_X(x) < \infty.$$

*If X is integrable, then its **expected value** is defined as*

$$\mathbb{E}(X) \doteq \sum_{x \in \mathscr{S}_X} x p_X(x).$$

*A continuous random variable X with p.d.f. f_X is **integrable** if*

$$\int_{-\infty}^{\infty} |x| f_X(x)\, dx < \infty.$$

*If X is integrable, then its **expected value** is defined as*

$$\mathbb{E}(X) \doteq \int_{-\infty}^{\infty} x f_X(x)\, dx.$$

In Definition 1.10 above, the expected value of a random variable can be interpreted as a weighted average of the possible values that the random variable can take on. The more 'likely' an outcome is, the more heavily it is weighted.

Example 1.11. *We say that $X \sim Cauchy(0,1)$ if its p.d.f. is given by*

$$f_X(x) \doteq \frac{1}{\pi(1+x^2)}, \qquad x \in \mathbb{R}.$$

The p.d.f. of the Cauchy distribution resembles that of the normal distribution, but it goes to 0 less quickly as $x \to \pm\infty$. A consequence of this is that X is not integrable, so its expected value is not defined. In order to see this, note that

$$\int_{-\infty}^{\infty} |x| f_X(x)\,dx \doteq \lim_{a,b\to\infty} \int_{-a}^{b} |x| f_X(x)\,dx,$$

and observe that for $a, b > 0$,

$$\begin{aligned}\int_{-a}^{b} |x| f_X(x)\,dx &= \int_{-a}^{b} \frac{|x|}{\pi(1+x^2)}\,dx \\ &= \int_{-a}^{0} \frac{|x|}{\pi(1+x^2)}\,dx + \int_{0}^{b} \frac{|x|}{\pi(1+x^2)}\,dx \\ &= \int_{0}^{a} \frac{x}{\pi(1+x^2)}\,dx + \int_{0}^{b} \frac{x}{\pi(1+x^2)}\,dx\end{aligned}$$

Using the change of variable $u \doteq x^2, du \doteq 2x\,dx$, we can evaluate

$$\begin{aligned}\int_{0}^{a} \frac{x}{\pi(1+x^2)}\,dx &= \frac{1}{2\pi}\int_{0}^{a^2} \frac{1}{1+u}\,du \\ &= \frac{1}{2\pi}\log(1+u)\Big|_{0}^{a^2} \\ &= \frac{1}{2\pi}\left(\log(1+a^2) - \log(1+0)\right) \\ &= \frac{1}{2\pi}\log(1+a^2),\end{aligned}$$

so it follows that

$$\int_{0}^{\infty} \frac{x}{\pi(1+x^2)}\,dx \doteq \lim_{a\to\infty} \int_{0}^{a} \frac{x}{\pi(1+x^2)}\,dx = \lim_{a\to\infty} \frac{1}{2\pi}\log(1+a^2) = \infty.$$

Similarly,

$$\int_{-\infty}^{0} \frac{x}{\pi(1+x^2)}\,dx = \infty,$$

so X is not integrable.

Nearly every random variable we study in this course will be integrable, and so will have a well-defined expected value. Accordingly, in the rest of this section it will be assumed that all random variables are integrable. The following result will be useful throughout this course.

Theorem 1.12. *Let $g : \mathbb{R} \to \mathbb{R}$ be a function.*

If X is a discrete random variable with p.m.f. p_X and state space $\mathcal{S}_X$, then

$$\mathbb{E}(g(X)) = \sum_{x\in\mathcal{S}_X} g(x)\,p_X(x).$$

If X is a continuous random variable with p.d.f. f_X, then

$$\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x) dx.$$

Theorem 1.12 tells us exactly how to calculate the expected value of a function of a random variable. For example, it allows us to compute the moments of X.

Definition 1.13. *The n-**th moment** ($n \in \mathbb{N}$) of a random variable X is the quantity $\mathbb{E}(X^n)$.*

*The **variance** of a random variable is defined as*

$$Var(X) \doteq \mathbb{E}((X - \mathbb{E}(X))^2) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2.$$

*The **covariance** of two random variables X and Y is defined as*

$$Cov(X, Y) \doteq \mathbb{E}\left[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\right] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

Note that in order to calculate $Cov(X, Y)$, we need to know the joint distribution *of X and Y. We will come to this shortly.*

Remark 1.14. *In order to calculate the n-th moment of a random variable, we take the function g in Theorem* 1.12 *to be $g(x) \doteq x^n$.*

In order to calculate the variance of X, we can take $g(x) \doteq (x - \mathbb{E}(X))^2$ in Theorem 1.12.

Some important properties of expectation are below. Here X and Y are random variables and $a, b \in \mathbb{R}$ are (non-random) constants.

(1) $\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$ *(linearity)*
(2) If $\mathbb{P}(X \leq Y) = 1$, *then* $\mathbb{E}(X) \leq \mathbb{E}(Y)$ *(monotonicity)*
(3) $\mathbb{E}(X + a) = \mathbb{E}(X) + a$
(4) $Var(aX) = a^2 Var(X)$
(5) $Var(X + a) = Var(X)$
(6) $Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$.
(7) If X and Y are independent, then $Var(X + Y) = Var(X) + Var(Y)$.

The following example shows how to calculate the expected value and variance of an exponential random variable.

Example 1.15. *Let $X \sim Exp(\lambda)$. Then*

$$\begin{aligned}\mathbb{E}(X) &= \int_{-\infty}^{\infty} x f_X(x) dx \\ &= \int_0^{\infty} x \cdot \lambda e^{-\lambda x} dx\end{aligned}$$

Integrating by parts with $u \doteq \lambda x$, $dv \doteq e^{-\lambda x}$, *so that* $du = \lambda dx$, $v = -\frac{1}{\lambda}e^{-\lambda x}$, *we obtain*

$$\begin{aligned}
\mathbb{E}(X) &= \int_0^\infty x \cdot \lambda e^{-\lambda x} dx \\
&= \int_0^\infty u dv \\
&= uv\Big|_0^\infty - \int_0^\infty v du \\
&= -xe^{-\lambda x}\Big|_0^\infty + \int_0^\infty e^{-\lambda x} dx \\
&= 0 + \left(-\frac{1}{\lambda}e^{-\lambda x}\Big|_0^\infty\right) \\
&= \frac{1}{\lambda}.
\end{aligned}$$

The second moment of X *can be calculated by integrating by parts with* $u \doteq \lambda x^2$ *and* $dv \doteq e^{-\lambda x}$. *Then we have* $du = 2\lambda x dx$, $v \doteq -\frac{1}{\lambda}e^{-\lambda x}$, *so that*

$$\begin{aligned}
\mathbb{E}(X^2) &= \int_0^\infty x^2 f_X(x) dx \\
&= \int_0^\infty \lambda x^2 e^{-\lambda x} dx \\
&= uv\Big|_0^\infty - \int_0^\infty v du \\
&= \frac{2}{\lambda^2}.
\end{aligned}$$

It follows that

$$Var(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}.$$

Below we derive the expected value and variance of a geometric random variable.

Example 1.16. *Let* $X \sim Geometric(p)$ *for some* $p \in (0,1)$, *so that the p.m.f. of* X *is given by*

$$p_X(x) = (1-p)^{x-1}p, \qquad x \in \mathbb{N}.$$

Calculate $\mathbb{E}(X)$ *and* $Var(X)$.

We begin by recalling that if $|r| < 1$, *then*

$$h(r) \doteq \sum_{x=1}^{\infty} r^x = \frac{1}{1-r}.$$

Therefore

$$h'(r) = \sum_{x=1}^{\infty} x r^{x-1} = \frac{d}{dr}\left(\frac{1}{1-r}\right) = \frac{1}{(1-r)^2},$$

and

$$h''(r) = \sum_{x=1}^{\infty} x(x-1) r^{x-2} = \frac{2}{(1-r)^3}.$$

Note that

$$\begin{aligned}\mathbb{E}(X) &= \sum_{x=1}^{\infty} x p_X(x) \\ &= p\sum_{x=1}^{\infty} x(1-p)^{x-1} \\ &= ph'(1-p) \\ &= p\frac{1}{(1-(1-p))^2} \\ &= \frac{1}{p}.\end{aligned}$$

Similarly,

$$\begin{aligned}\mathbb{E}(X^2) &= \sum_{x=1}^{\infty} x^2 p_X(x) \\ &= p\sum_{x=1}^{\infty} x^2(1-p)^{x-1} \\ &= p\sum_{x=1}^{\infty} (x^2 - x + x)(1-p)^{x-1} \\ &= p(1-p)\sum_{x=1}^{\infty} x(x-1)(1-p)^{x-2} + p\sum_{x=1}^{\infty} x(1-p)^{x-1} \\ &= p(1-p)h''(1-p) + ph'(1-p) \\ &= \frac{2p(1-p)}{p^3} + \frac{p}{p^2} \\ &= \frac{2-p}{p^2}.\end{aligned}$$

It follows that

$$Var(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = \frac{2-p}{p^2} - \frac{1}{p^2} = \frac{1-p}{p^2}.$$

1.6. **Joint Distribution of Random Variables.** The joint distribution of two random variables X and Y describes how the two random variables behave when they are viewed *together*. For an example of the subleties that can arise with joint distributions, let $X \sim \mathcal{N}(0,1)$ be a standard normal random variable, and let $Y \doteq -X$. Then $Y \sim \mathcal{N}(0,1)$ as well, so

$$\mathbb{P}(X \geq 0) = \mathbb{P}(Y \geq 0) = 1/2.$$

However,

$$\mathbb{P}(X \geq 0, Y \geq 0) = \mathbb{P}(X \geq 0, -X \geq 0) = 0,$$

so when we view X and Y together they behave very differently than they do on their own.

To understand the joint distribution of two random variables, we begin by introducing the notion of a bivariate random variable. A **bivariate random variable** on some probability space $(\Omega, \mathscr{F}, \mathbb{P})$ is a random variable of the form $(X,Y) : \Omega \to \mathbb{R}^2$, meaning that for each $\omega \in \Omega$, there is some $(x,y) \in \mathbb{R}^2$ such that $(X(\omega), Y(\omega)) = (x,y) \in \mathbb{R}^2$.

If X and Y are jointly discrete, then there is some p.m.f. $p_{(X,Y)} : \mathbb{R} \to [0,1]$ such that for each pair of events $A, B \subseteq \mathbb{R}$,

$$\mathbb{P}((X,Y) \in A \times B) = \sum_{x,y \in A\times B} p_{(X,Y)}(x,y) \doteq \sum_{x\in A}\sum_{y\in B} p_{(X,Y)}(x,y)$$

We refer to $p_{(X,Y)}$ as the **joint p.m.f.** of (X,Y).

Similarly, if (X,Y) are jointly continuous, then there is a **joint p.d.f.** $f_{X,Y}:\mathbb{R}\to\mathbb{R}_+$ such that for each pair of events $A,B\subseteq\mathbb{R}$,

$$P((X,Y)\in A\times B)=\int_{A\times B}f_{(X,Y)}(x,y)d(x,y)=\int_A\int_B f_{(X,Y)}(x,y)dydx=\int_B\int_A f_{(X,Y)}(x,y)dxdy.$$

Note that we can exchange the order of integration above due to Fubini's theorem. [2]

The joint p.m.f. and joint p.d.f. also give rise to a joint c.d.f. , which, as in the univariate case, fully characterizes the joint distribution of (X,Y). The **joint c.d.f.** of (X,Y) is the function $F_{(X,Y)}:\mathbb{R}^2\to[0,1]$ defined as

$$F_{(X,Y)}(x,y)\doteq\mathbb{P}(X\le x,Y\le y).$$

Note that if X and Y are jointly discrete with joint p.m.f. $p_{(X,Y)}$, then their joint c.d.f. is given by

$$F_{(X,Y)}(x,y)\doteq\mathbb{P}(X\le x,Y\le y)=\sum_{i\le x}\sum_{j\le y}p_{(X,Y)}(i,j).$$

Similarly, if X and Y are jointly continuous with joint p.d.f. $f_{(X,Y)}$, then their joint c.d.f. is given by

$$\begin{aligned}F_{(X,Y)}(x,y)&\doteq\mathbb{P}(X\le x,Y\le y)\\&=\int_{(-\infty,x]\times(-\infty,y]}f_{(X,Y)}(u,v)d(u,v)\\&=\int_{-\infty}^x\int_{-\infty}^y f_{(X,Y)}(u,v)dvdu\\&=\int_{-\infty}^y\int_{-\infty}^x f_{(X,Y)}(u,v)dudv,\end{aligned}$$

where we once more use Fubini's theorem to justify the equivalences.

Proposition 1.17. *If we know the joint distribution of (X,Y), we can also recover their individual (i.e., marginal) distributions. In the discrete case we have*

$$p_X(x)=\sum_{y\in\mathscr{S}_Y}p_{(X,Y)}(x,y),\qquad p_Y(y)=\sum_{x\in\mathscr{S}_X}p_{(X,Y)}(x,y).$$

and in the continuous case we have

$$f_X(x)=\int_{-\infty}^{\infty}f_{(X,Y)}(x,y)dy,\qquad f_Y(y)=\int_{-\infty}^{\infty}f_{(X,Y)}(x,y)dx.$$

Additionally, recall that two discrete random variables X and Y are independent if and only if for all $x\in\mathscr{S}_X$, $y\in\mathscr{S}_Y$ we have

$$p_{(X,Y)}(x,y)=p_X(x)p_Y(y).$$

Similarly, two continuous random variables X and Y are independent if and only if for all $x,y\in\mathbb{R}$ we have

$$f_{(X,Y)}(x,y)=f_X(x)f_Y(y).$$

Recall from Definition 1.13 that in order to calculate the covariance of two random variables X and Y, we need to calculate $\mathbb{E}(XY)$. The following formula allows us to compute this as well as other quantities such as $\mathbb{E}(X^Y)$, $\mathbb{E}(X^2Y^2)$, and so on.

Let $g:\mathbb{R}^2\to\mathbb{R}$ be a function. If (X,Y) is jointly discrete with joint p.m.f. $p_{(X,Y)}$, then

$$\mathbb{E}(g(X,Y))=\sum_{(x,y)\in\mathscr{S}_X\times\mathscr{S}_Y}g(x,y)p_{(X,Y)}(x,y)=\sum_{x\in\mathscr{S}_X}\sum_{y\in\mathscr{S}_Y}g(x,y)p_{(X,Y)}(x,y).$$

Similarly, if (X,Y) is jointly continuous with joint p.d.f. $f_{(X,Y)}$, then

$$\mathbb{E}(g(X,Y))=\int_{\mathbb{R}^2}g(x,y)f_{(X,Y)}(x,y)d(x,y)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}g(x,y)f_{(X,Y)}(x,y)dxdy=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}g(x,y)f_{(X,Y)}(x,y)dydx.$$

[2] See `https://en.wikipedia.org/wiki/Fubini's_theorem`

If we take $g(x,y) \doteq xy$, then in the discrete case this yields

$$\mathbb{E}(XY) = \sum_{x\in\mathscr{S}_X}\sum_{y\in\mathscr{S}_Y} xy p_{(X,Y)}(x,y).$$

In the continuous case we have

$$\mathbb{E}(XY) = \int_{\mathbb{R}^2} xy f_{(X,Y)}(x,y)d(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy f_{(X,Y)}(x,y)dxdy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy f_{(X,Y)}dydx.$$

Using this, if we know the joint distribution of X and Y, then we can calculate their covariance:

$$\mathrm{Cov}(X,Y) = \mathbb{E}((X-\mathbb{E}(X))(Y-\mathbb{E}(Y))) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

In the following example we use the joint p.d.f. of a pair of random variables to calculate their covariance.

Example 1.18. *Let X and Y be continuous random variables with joint p.d.f.*

$$f_{(X,Y)}(x,y) \doteq 3x, \qquad 0 \le y \le x \le 1.$$

Calculate the marginal densities of X and Y. Determine whether X and Y are independent and calculate $Cov(X,Y)$.

The marginal density of X is given by

$$f_X(x) = \int_{-\infty}^{\infty} f_{(X,Y)}(x,y)dy = \int_0^x 3x dy = 3x^2, \qquad 0 \le x \le 1,$$

and the marginal density of Y is given by

$$f_Y(y) = \int_{-\infty}^{\infty} f_{(X,Y)}(x,y)dx = \int_y^1 3x dx = \frac{3x^2}{2}\Big|_y^1 = \frac{3(1-y^2)}{2}, \qquad 0 \le y \le 1.$$

Since the joint density $f_{(X,Y)}$ is not the product of the marginal densities f_X and f_Y, it follows that X and Y are not independent. Therefore

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x)dx = \int_0^1 x\cdot 3x^2 dx = \frac{3}{4},$$

and

$$\mathbb{E}(Y) = \int_{-\infty}^{\infty} y f_Y(y)dy = \int_0^1 y\frac{3(1-y^2)}{2}dy = \frac{3}{8}.$$

Finally,

$$\begin{aligned}
\mathbb{E}(XY) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy f_{(X,Y)}(x,y)dydx \\
&= \int_0^1\int_0^x xy\cdot 3x dydx \\
&= \int_0^1 3x^2\left(\int_0^x y dy\right)dx \\
&= \int_0^1 \frac{x^2}{2}\cdot 3x^2 dx \\
&= \frac{3}{10}.
\end{aligned}$$

Therefore

$$Cov(XY) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \frac{3}{10} - \frac{3}{4}\times\frac{3}{8} = \frac{3}{160}.$$

1.7. **Conditional Probability and Independence.** Conditional probability allows us to understand the how the outcome of one event (i.e., did the event happen or not) affects the outcome of another event. For instance, given two events A and B, it allows us to calculate the probability that A happens given that B has already happened (or will definitely happen, depending on the situation).

Definition 1.19. *Given two events A and B, the conditional probability of A given B is defined as*

$$\mathbb{P}(A|B) \doteq \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)},$$

as long as $\mathbb{P}(B) > 0$.

Conditional probability has similar properties to regular probability, namely for events A, B, and C, the following hold whenever $\mathbb{P}(B) > 0$:

(1) $\mathbb{P}(A|B) = 1 - \mathbb{P}(A^c|B)$
(2) $\mathbb{P}(A \cup C|B) = \mathbb{P}(A|B) + \mathbb{P}(C|B) - \mathbb{P}(A \cap C|B)$
(3) If $A \subseteq C$, then $\mathbb{P}(A|B) \leq \mathbb{P}(C|B)$
(4) $\mathbb{P}(\emptyset|B) = 0$

The following proposition illustrates how one of the many uses of conditional probability. Recall that a **partition** of the sample space Ω is a (finite or infinite) collection of pairwise disjoint (i.e., $B_i \cap B_j = \emptyset$ whenever $i \neq j$) sets $\{B_1, B_2, \ldots\}$ such that $\bigcup_{n=1}^{\infty} B_n = \Omega$ and $\mathbb{P}(B_i) > 0$ for all $i \geq \in \mathbb{N}$.

Proposition 1.20. *(Law of Total Probability) Let $\{B_1, B_2, \ldots\}$ be a partition of Ω. Then for each event $A \subseteq \Omega$,*

$$\mathbb{P}(A) = \sum_{i \geq 1} \mathbb{P}(A|B_i)\mathbb{P}(B_i).$$

A simple consequence of this law is that for any events A, B satisfying $\mathbb{P}(B) > 0$, we have

$$\mathbb{P}(A) = \mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c).$$

The following example illustrates how one can apply the Law of Total Probability to simplify some calculations.

Example 1.21. *You have three jars of marbles. Jar 1 contains 75 red and 25 blue marbles, jar 2 contains 60 red and 40 blue marbles, and jar 3 contains 45 red and 55 blue marbles. You choose one fo the jars at random then randomly draw a marble from that jar. What is the probability that you draw a red marble?*

Let A denote the event that you draw a red marble, and for $i \in \{1,2,3\}$, let B_i denote the event that you choose jar i. Then

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}(A|B_1)\mathbb{P}(B_1) + \mathbb{P}(A|B_2)\mathbb{P}(B_2) + \mathbb{P}(A|B_3)\mathbb{P}(B_3) \\ &= \frac{1}{3}\left(\frac{75}{100} + \frac{60}{100} + \frac{45}{100}\right) \\ &= \frac{3}{5}\end{aligned}$$

The following result is known as Bayes' formula. It follows immediately from the definition of conditional probability.

Proposition 1.22. *(Bayes' Formula) For any events A and B with $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$, we have*

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)}.$$

The following example illustrates how one can apply Bayes' formula.

Example 1.23. *You have a bag with 100 coins. Of these coins, 99 are real coins, namely they are fair and have heads on one side and tails on the other, but one of the coins has heads on both sides.*

You pick a coin at random, and do not check whether it is real or fake. You flip the coin $n \in \mathbb{N}$ times in a row, and it lands on heads all n times. What is the probability that you picked the fake *coin?*

Let A be the event that you picked a fake coin, and let B be the event that a coin lands on heads for all n flips. Then

$$\begin{aligned}\mathbb{P}(B) &= \mathbb{P}(B|A)\mathbb{P}(A) + \mathbb{P}(B|A^c)\mathbb{P}(A^c) \\ &= 1 \cdot \frac{1}{100} + \left(\frac{1}{2}\right)^n \cdot \frac{99}{100} \\ &= \frac{1}{100}\left(1 + 99\left(\frac{1}{2}\right)^n\right).\end{aligned}$$

Consequently,

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\frac{1}{100} \cdot 1}{\frac{1}{100}\left(1 + 99\left(\frac{1}{2}\right)^n\right)} = \frac{1}{1 + 99\left(\frac{1}{2}\right)^n}.$$

Since $\mathbb{P}(A|B)$ increases towards 1 as $n \to \infty$, it follows that the larger n is, the more likely it is that you drew the fake coin.

Finally, we recall the notion of independent random variables once more.

Definition 1.24. *Two random variables X and Y are **independent** if and only if for all events A and B,*

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B).$$

*Similarly, random variables $X_1, \ldots, X_n$ are **independent** if and only if for all events $A_1, \ldots, A_n$, we have*

$$\mathbb{P}(X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \in A_i) \doteq \mathbb{P}(X_1 \in A_1)\mathbb{P}(X_2 \in A_2) \cdots \mathbb{P}(X_n \in A_n).$$

The following proposition gives several different criteria that can be used to check for independence of random variables. It is stated only for a pair of random variables, but the analogous result holds for a collection of $n > 2$ random variables as well.

Proposition 1.25. *Let X and Y be random variables with c.d.f.'s F_X and F_Y, respectively. Let $F_{(X,Y)}$ denote the joint c.d.f. of (X, Y). Then the following are equivalent:*

(1) X and Y are independent.
(2) For all $(x, y) \in \mathbb{R}^2$, $F_{(X,Y)}(x, y) = F_X(x)F_Y(y)$.
(3) If X and Y are discrete, then for all $(x, y) \in \mathscr{S}_X \times \mathscr{S}_y$, $p_{(X,Y)}(x, y) = p_X(x)p_Y(y)$. Similarly, if X and Y are continuous, then for all $(x, y) \in \mathbb{R}^2$, $f_{(X,Y)}(x, y) = f_X(x)f_Y(y)$.

Additionally, if X and Y are independent, then $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.

This brings us to the end of our review of the material from PSTAT 120A. If you are feeling uncomfortable with any of this material, please let me know as soon as possible, as these formulas, definitions, and ideas will be used regularly throughout this course.