2

Discrete Random Variables

Contents

2.1 BASIC CONCEPTS

In many probabilistic models, the outcomes are of a numerical nature, e.g., if they correspond to instrument readings or stock prices. In other experiments, the outcomes are not numerical, but they may be associated with some numerical values of interest. For example, if the experiment is the selection of students from a given population, we may wish to consider their grade point average. When dealing with such numerical values, it is often useful to assign probabilities to them. This is done through the notion of a **random variable**, the focus of the present chapter.

Given an experiment and the corresponding set of possible outcomes (the sample space), a random variable associates a particular number with each outcome; see Fig. 2.1. We refer to this number as the **numerical value** or the **experimental value** of the random variable. Mathematically, **a random variable is a real-valued function of the experimental outcome**.

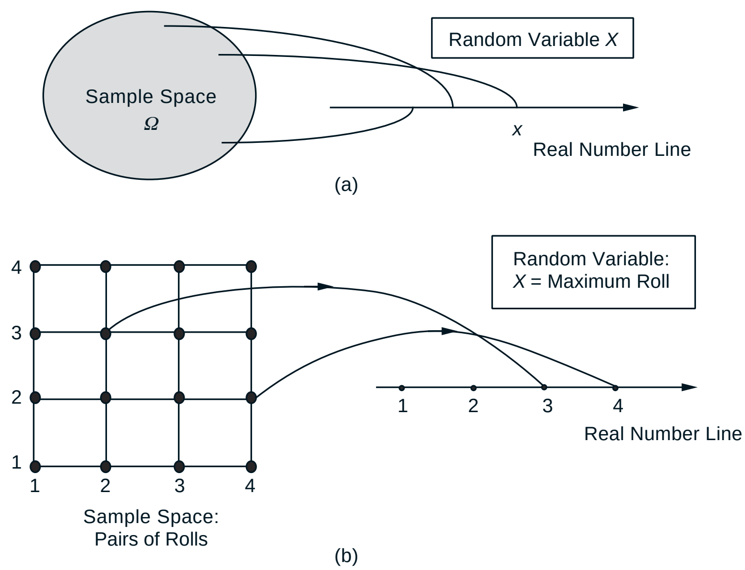

Figure 2.1: (a) Visualization of a random variable. It is a function that assigns a numerical value to each possible outcome of the experiment. (b) An example of a random variable. The experiment consists of two rolls of a 4-sided die, and the random variable is the maximum of the two rolls. If the outcome of the experiment is $(4, 2)$, the experimental value of this random variable is 4.

Here are some examples of random variables:

(a) In an experiment involving a sequence of 5 tosses of a coin, the number of heads in the sequence is a random variable. However, the 5-long sequence of heads and tails is not considered a random variable because it does not have an explicit numerical value.

(b) In an experiment involving two rolls of a die, the following are examples of random variables:

(1) The sum of the two rolls.

(2) The number of sixes in the two rolls.

(3) The second roll raised to the fifth power.

(c) In an experiment involving the transmission of a message, the time needed to transmit the message, the number of symbols received in error, and the delay with which the message is received are all random variables.

There are several basic concepts associated with random variables, which are summarized below.

Main Concepts Related to Random Variables

Starting with a probabilistic model of an experiment:

- A **random variable** is a real-valued function of the outcome of the experiment.
- A **function of a random variable** defines another random variable.
- We can associate with each random variable certain "averages" of interest, such the **mean** and the **variance**.
- A random variable can be **conditioned** on an event or on another random variable.
- There is a notion of **independence** of a random variable from an event or from another random variable.

A random variable is called **discrete** if its **range** (the set of values that it can take) is finite or at most countably infinite. For example, the random variables mentioned in (a) and (b) above can take at most a finite number of numerical values, and are therefore discrete.

A random variable that can take an uncountably infinite number of values is not discrete. For an example, consider the experiment of choosing a point a from the interval $[-1, 1]$. The random variable that associates the numerical value a^2 to the outcome a is not discrete. On the other hand, the random variable that associates with a the numerical value

$$\mathrm{sgn}(a) = \begin{cases} 1 & \text{if } a > 0, \\ 0 & \text{if } a = 0, \\ -1 & \text{if } a < 0, \end{cases}$$

is discrete.

In this chapter, we focus exclusively on discrete random variables, even though we will typically omit the qualifier "discrete."

> **Concepts Related to Discrete Random Variables**
>
> Starting with a probabilistic model of an experiment:
>
> - A **discrete random variable** is a real-valued function of the outcome of the experiment that can take a finite or countably infinite number of values.
> - A (discrete) random variable has an associated **probability mass function** (PMF), which gives the probability of each numerical value that the random variable can take.
> - A **function of a random variable** defines another random variable, whose PMF can be obtained from the PMF of the original random variable.

We will discuss each of the above concepts and the associated methodology in the following sections. In addition, we will provide examples of some important and frequently encountered random variables. In Chapter 3, we will discuss general (not necessarily discrete) random variables.

Even though this chapter may appear to be covering a lot of new ground, this is not really the case. The general line of development is to simply take the concepts from Chapter 1 (probabilities, conditioning, independence, etc.) and apply them to random variables rather than events, together with some appropriate new notation. The only genuinely new concepts relate to means and variances.

2.2 PROBABILITY MASS FUNCTIONS

The most important way to characterize a random variable is through the probabilities of the values that it can take. For a discrete random variable X, these are captured by the **probability mass function** (PMF for short) of X, denoted p_X. In particular, if x is any possible value of X, the **probability mass** of x, denoted $p_X(x)$, is the probability of the event $\{X = x\}$ consisting of all outcomes that give rise to a value of X equal to x:

$$p_X(x) = \mathbf{P}\big(\{X = x\}\big).$$

For example, let the experiment consist of two independent tosses of a fair coin, and let X be the number of heads obtained. Then the PMF of X is

$$p_X(x) = \begin{cases} 1/4 & \text{if } x = 0 \text{ or } x = 2, \\ 1/2 & \text{if } x = 1, \\ 0 & \text{otherwise.} \end{cases}$$

In what follows, we will often omit the braces from the event/set notation, when no ambiguity can arise. In particular, we will usually write $\mathbf{P}(X = x)$ in place of the more correct notation $\mathbf{P}\big(\{X = x\}\big)$. We will also adhere to the following convention throughout: **we will use upper case characters to denote random variables, and lower case characters to denote real numbers such as the numerical values of a random variable.**

Note that

$$\sum_x p_X(x) = 1,$$

where in the summation above, x ranges over all the possible numerical values of X. This follows from the additivity and normalization axioms, because the events $\{X = x\}$ are disjoint and form a partition of the sample space, as x ranges over all possible values of X. By a similar argument, for any set S of real numbers, we also have

$$\mathbf{P}(X \in S) = \sum_{x \in S} p_X(x).$$

For example, if X is the number of heads obtained in two independent tosses of a fair coin, as above, the probability of at least one head is

$$\mathbf{P}(X > 0) = \sum_{x>0} p_X(x) = \frac{1}{2} + \frac{1}{4} = \frac{3}{4}.$$

Calculating the PMF of X is conceptually straightforward, and is illustrated in Fig. 2.2.

Calculation of the PMF of a Random Variable X

For each possible value x of X:

1. Collect all the possible outcomes that give rise to the event $\{X = x\}$.
2. Add their probabilities to obtain $p_X(x)$.

The Bernoulli Random Variable

Consider the toss of a biased coin, which comes up a head with probability p, and a tail with probability $1-p$. The **Bernoulli** random variable takes the two values 1 and 0, depending on whether the outcome is a head or a tail:

$$X = \begin{cases} 1 & \text{if a head,} \\ 0 & \text{if a tail.} \end{cases}$$

Its PMF is

$$p_X(x) = \begin{cases} p & \text{if } x = 1, \\ 1-p & \text{if } x = 0. \end{cases}$$

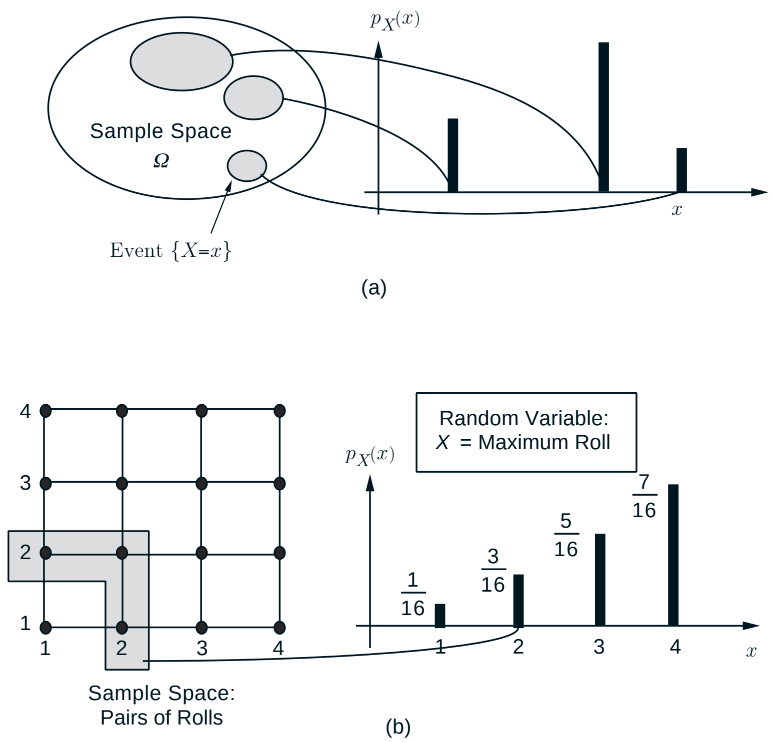

Figure 2.2: (a) Illustration of the method to calculate the PMF of a random variable X. For each possible value x, we collect all the outcomes that give rise to $X = x$ and add their probabilities to obtain $p_X(x)$. (b) Calculation of the PMF p_X of the random variable X = maximum roll in two independent rolls of a fair 4-sided die. There are four possible values x, namely, 1, 2, 3, 4. To calculate $p_X(x)$ for a given x, we add the probabilities of the outcomes that give rise to x. For example, there are three outcomes that give rise to $x = 2$, namely, $(1, 2), (2, 2), (2, 1)$. Each of these outcomes has probability $1/16$, so $p_X(2) = 3/16$, as indicated in the figure.

For all its simplicity, the Bernoulli random variable is very important. In practice, it is used to model generic probabilistic situations with just two outcomes, such as:

(a) The state of a telephone at a given time that can be either free or busy.

(b) A person who can be either healthy or sick with a certain disease.

(c) The preference of a person who can be either for or against a certain political candidate.

Furthermore, by combining multiple Bernoulli random variables, one can construct more complicated random variables.

The Binomial Random Variable

A biased coin is tossed n times. At each toss, the coin comes up a head with probability p, and a tail with probability $1-p$, independently of prior tosses. Let X be the number of heads in the n-toss sequence. We refer to X as a **binomial** random variable **with parameters** n **and** p. The PMF of X consists of the binomial probabilities that were calculated in Section 1.4:

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

(Note that here and elsewhere, we simplify notation and use k, instead of x, to denote the experimental values of integer-valued random variables.) The normalization property $\sum_x p_X(x) = 1$, specialized to the binomial random variable, is written as

$$\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = 1.$$

Some special cases of the binomial PMF are sketched in Fig. 2.3.

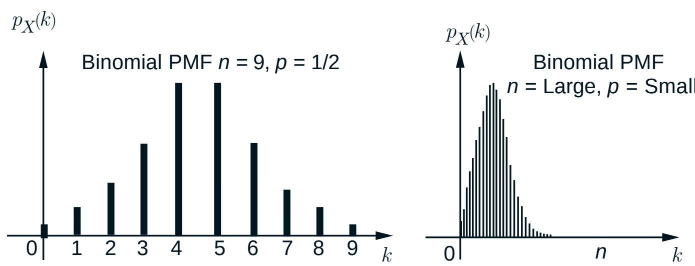

Figure 2.3: The PMF of a binomial random variable. If $p = 1/2$, the PMF is symmetric around $n/2$. Otherwise, the PMF is skewed towards 0 if $p < 1/2$, and towards n if $p > 1/2$.

The Geometric Random Variable

Suppose that we repeatedly and independently toss a biased coin with probability of a head p, where $0 < p < 1$. The **geometric** random variable is the number X of tosses needed for a head to come up for the first time. Its PMF is given by

$$p_X(k) = (1-p)^{k-1} p, \qquad k = 1, 2, \ldots,$$

since $(1-p)^{k-1}p$ is the probability of the sequence consisting of $k-1$ successive tails followed by a head; see Fig. 2.4. This is a legitimate PMF because

$$\sum_{k=1}^{\infty} p_X(k) = \sum_{k=1}^{\infty} (1-p)^{k-1} p = p \sum_{k=0}^{\infty} (1-p)^k = p \cdot \frac{1}{1-(1-p)} = 1.$$

Naturally, the use of coin tosses here is just to provide insight. More generally, we can interpret the geometric random variable in terms of repeated independent trials until the first "success." Each trial has probability of success p and the number of trials until (and including) the first success is modeled by the geometric random variable.

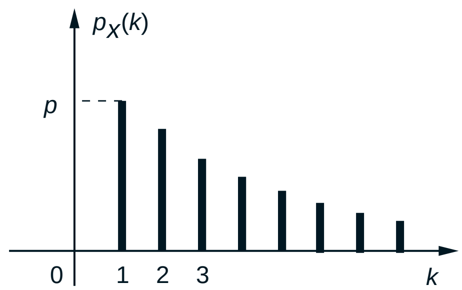

Figure 2.4: The PMF

$$p_X(k) = (1-p)^{k-1}p, \qquad k = 1, 2, \ldots,$$

of a geometric random variable. It decreases as a geometric progression with parameter $1-p$.

The Poisson Random Variable

A Poisson random variable takes nonnegative integer values. Its PMF is given by

$$p_X(k) = e^{-\lambda}\frac{\lambda^k}{k!}, \qquad k = 0, 1, 2, \ldots,$$

where λ is a positive parameter characterizing the PMF, see Fig. 2.5. It is a legitimate PMF because

$$\sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k}{k!} = e^{-\lambda}\left(1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \cdots\right) = e^{-\lambda}e^{\lambda} = 1.$$

To get a feel for the Poisson random variable, think of a binomial random variable with very small p and very large n. For example, consider the number of typos in a book with a total of n words, when the probability p that any one word is misspelled is very small (associate a word with a coin toss which comes a head when the word is misspelled), or the number of cars involved in accidents in a city on a given day (associate a car with a coin toss which comes a head when the car has an accident). Such a random variable can be well-modeled as a Poisson random variable.

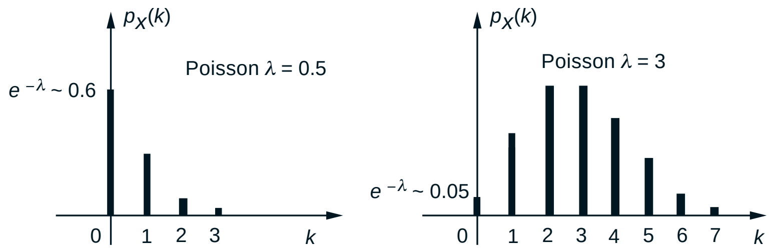

Figure 2.5: The PMF $e^{-\lambda}\frac{\lambda^k}{k!}$ of the Poisson random variable for different values of λ. Note that if $\lambda < 1$, then the PMF is monotonically decreasing, while if $\lambda > 1$, the PMF first increases and then decreases as the value of k increases (this is shown in the end-of-chapter problems).

More precisely, the Poisson PMF with parameter λ is a good approximation for a binomial PMF with parameters n and p, provided $\lambda = np$, n is very large, and p is very small, i.e.,

$$e^{-\lambda}\frac{\lambda^k}{k!} \approx \frac{n!}{(n-k)!k!}p^k(1-p)^{n-k}, \qquad k = 0, 1, \ldots, n.$$

In this case, using the Poisson PMF may result in simpler models and calculations. For example, let $n = 100$ and $p = 0.01$. Then the probability of $k = 5$ successes in $n = 100$ trials is calculated using the binomial PMF as

$$\frac{100!}{95!\,5!}0.01^5(1-0.01)^{95} = 0.00290.$$

Using the Poisson PMF with $\lambda = np = 100 \cdot 0.01 = 1$, this probability is approximated by

$$e^{-1}\frac{1}{5!} = 0.00306.$$

We provide a formal justification of the Poisson approximation property in the end-of-chapter problems and also in Chapter 5, where we will further interpret it, extend it, and use it in the context of the Poisson process.

2.3 FUNCTIONS OF RANDOM VARIABLES

Consider a probability model of today's weather, let the random variable X be the temperature in degrees Celsius, and consider the transformation $Y = 1.8X + 32$, which gives the temperature in degrees Fahrenheit. In this example, Y is a **linear** function of X, of the form

$$Y = g(X) = aX + b,$$

where a and b are scalars. We may also consider nonlinear functions of the general form

$$Y = g(X).$$

For example, if we wish to display temperatures on a logarithmic scale, we would want to use the function $g(X) = \log X$.

If $Y = g(X)$ is a function of a random variable X, then Y is also a random variable, since it provides a numerical value for each possible outcome. This is because every outcome in the sample space defines a numerical value x for X and hence also the numerical value $y = g(x)$ for Y. If X is discrete with PMF p_X, then Y is also discrete, and its PMF p_Y can be calculated using the PMF of X. In particular, to obtain $p_Y(y)$ for any y, we add the probabilities of all values of x such that $g(x) = y$:

$$p_Y(y) = \sum_{\{x \mid g(x)=y\}} p_X(x).$$

Example 2.1. Let $Y = |X|$ and let us apply the preceding formula for the PMF p_Y to the case where

$$p_X(x) = \begin{cases} 1/9 & \text{if } x \text{ is an integer in the range } [-4,4], \\ 0 & \text{otherwise.} \end{cases}$$

The possible values of Y are $y = 0, 1, 2, 3, 4$. To compute $p_Y(y)$ for some given value y from this range, we must add $p_X(x)$ over all values x such that $|x| = y$. In particular, there is only one value of X that corresponds to $y = 0$, namely $x = 0$. Thus,

$$p_Y(0) = p_X(0) = \frac{1}{9}.$$

Also, there are two values of X that correspond to each $y = 1, 2, 3, 4$, so for example,

$$p_Y(1) = p_X(-1) + p_X(1) = \frac{2}{9}.$$

Thus, the PMF of Y is

$$p_Y(y) = \begin{cases} 2/9 & \text{if } y = 1,2,3,4, \\ 1/9 & \text{if } y = 0, \\ 0 & \text{otherwise.} \end{cases}$$

For another related example, let $Z = X^2$. To obtain the PMF of Z, we can view it either as the square of the random variable X or as the square of the random variable Y. By applying the formula $p_Z(z) = \sum_{\{x \mid x^2=z\}} p_X(x)$ or the formula $p_Z(z) = \sum_{\{y \mid y^2=z\}} p_Y(y)$, we obtain

$$p_Z(z) = \begin{cases} 2/9 & \text{if } z = 1,4,9,16, \\ 1/9 & \text{if } z = 0, \\ 0 & \text{otherwise.} \end{cases}$$

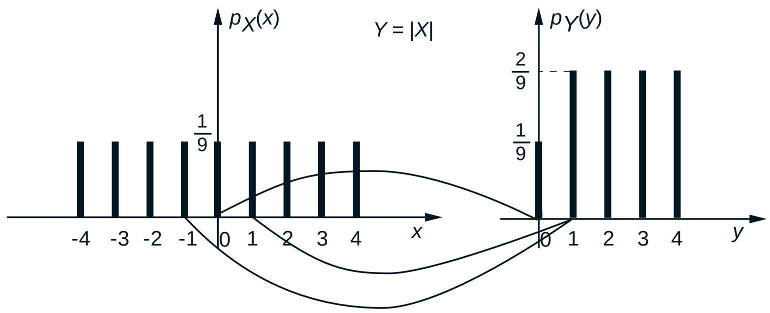

Figure 2.7: The PMFs of X and $Y = |X|$ in Example 2.1.

2.4 EXPECTATION, MEAN, AND VARIANCE

The PMF of a random variable X provides us with several numbers, the probabilities of all the possible values of X. It would be desirable to summarize this information in a single representative number. This is accomplished by the **expectation** of X, which is a weighted (in proportion to probabilities) average of the possible values of X.

As motivation, suppose you spin a wheel of fortune many times. At each spin, one of the numbers $m_1, m_2, \ldots, m_n$ comes up with corresponding probability $p_1, p_2, \ldots, p_n$, and this is your monetary reward from that spin. What is the amount of money that you "expect" to get "per spin"? The terms "expect" and "per spin" are a little ambiguous, but here is a reasonable interpretation.

Suppose that you spin the wheel k times, and that k_i is the number of times that the outcome is m_i. Then, the total amount received is $m_1k_1 + m_2k_2 + \cdots + m_nk_n$. The amount received per spin is

$$M = \frac{m_1k_1 + m_2k_2 + \cdots + m_nk_n}{k}.$$

If the number of spins k is very large, and if we are willing to interpret probabilities as relative frequencies, it is reasonable to anticipate that m_i comes up a fraction of times that is roughly equal to p_i:

$$p_i \approx \frac{k_i}{k}, \qquad i = 1, \ldots, n.$$

Thus, the amount of money per spin that you "expect" to receive is

$$M = \frac{m_1k_1 + m_2k_2 + \cdots + m_nk_n}{k} \approx m_1p_1 + m_2p_2 + \cdots + m_np_n.$$

Motivated by this example, we introduce an important definition.

Expectation

We define the **expected value** (also called the **expectation** or the **mean**) of a random variable X, with PMF $p_X(x)$, by†

$$\mathbf{E}[X] = \sum_x x p_X(x).$$

Example 2.2. Consider two independent coin tosses, each with a 3/4 probability of a head, and let X be the number of heads obtained. This is a binomial random variable with parameters $n = 2$ and $p = 3/4$. Its PMF is

$$p_X(k) = \begin{cases} (1/4)^2 & \text{if } k = 0, \\ 2 \cdot (1/4) \cdot (3/4) & \text{if } k = 1, \\ (3/4)^2 & \text{if } k = 2, \end{cases}$$

so the mean is

$$\mathbf{E}[X] = 0 \cdot \left(\frac{1}{4}\right)^2 + 1 \cdot \left(2 \cdot \frac{1}{4} \cdot \frac{3}{4}\right) + 2 \cdot \left(\frac{3}{4}\right)^2 = \frac{24}{16} = \frac{3}{2}.$$

It is useful to view the mean of X as a "representative" value of X, which lies somewhere in the middle of its range. We can make this statement more precise, by viewing the mean as the **center of gravity** of the PMF, in the sense explained in Fig. 2.8.

---

† When dealing with random variables that take a countably infinite number of values, one has to deal with the possibility that the infinite sum $\sum_x x p_X(x)$ is not well-defined. More concretely, we will say that the expectation is well-defined if $\sum_x |x| p_X(x) < \infty$. In that case, it is known that the infinite sum $\sum_x x p_X(x)$ converges to a finite value that is independent of the order in which the various terms are summed.

For an example where the expectation is not well-defined, consider a random variable X that takes the value 2^k with probability 2^{-k}, for $k = 1, 2, \ldots$. For a more subtle example, consider the random variable X that takes the values 2^k and -2^k with probability 2^{-k}, for $k = 2, 3, \ldots$. The expectation is again undefined, even though the PMF is symmetric around zero and one might be tempted to say that $\mathbf{E}[X]$ is zero.

Throughout this book, in lack of an indication to the contrary, we implicitly assume that the expected value of the random variables of interest is well-defined.

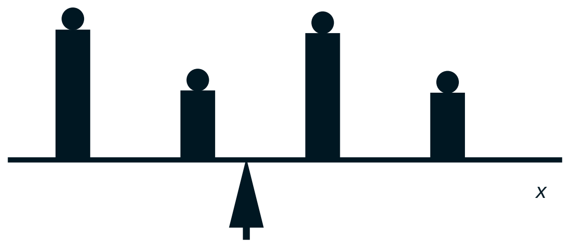

Figure 2.8: Interpretation of the mean as a center of gravity. Given a bar with a weight $p_X(x)$ placed at each point x with $p_X(x) > 0$, the center of gravity c is the point at which the sum of the torques from the weights to its left are equal to the sum of the torques from the weights to its right, that is,

$$\sum_x (x - c)p_X(x) = 0, \qquad \text{or} \quad c = \sum_x x p_X(x),$$

and the center of gravity is equal to the mean $\mathbf{E}[X]$.

There are many other quantities that can be associated with a random variable and its PMF. For example, we define the **2nd moment** of the random variable X as the expected value of the random variable X^2. More generally, we define the n**th moment** as $\mathbf{E}[X^n]$, the expected value of the random variable X^n. With this terminology, the 1st moment of X is just the mean.

The most important quantity associated with a random variable X, other than the mean, is its **variance**, which is denoted by $\text{var}(X)$ and is defined as the expected value of the random variable $\big(X - \mathbf{E}[X]\big)^2$, i.e.,

$$\text{var}(X) = \mathbf{E}\Big[\big(X - \mathbf{E}[X]\big)^2\Big].$$

Since $\big(X - \mathbf{E}[X]\big)^2$ can only take nonnegative values, the variance is always nonnegative.

The variance provides a measure of dispersion of X around its mean. Another measure of dispersion is the **standard deviation** of X, which is defined as the square root of the variance and is denoted by σ_X:

$$\sigma_X = \sqrt{\text{var}(X)}.$$

The standard deviation is often easier to interpret, because it has the same units as X. For example, if X measures length in meters, the units of variance are square meters, while the units of the standard deviation are meters.

One way to calculate $\text{var}(X)$, is to use the definition of expected value, after calculating the PMF of the random variable $\big(X - \mathbf{E}[X]\big)^2$. This latter random variable is a function of X, and its PMF can be obtained in the manner discussed in the preceding section.

Example 2.3. Consider the random variable X of Example 2.1, which has the PMF

$$p_X(x) = \begin{cases} 1/9 & \text{if } x \text{ is an integer in the range } [-4,4], \\ 0 & \text{otherwise.} \end{cases}$$

The mean $\mathbf{E}[X]$ is equal to 0. This can be seen from the symmetry of the PMF of X around 0, and can also be verified from the definition:

$$\mathbf{E}[X] = \sum_x x p_X(x) = \frac{1}{9} \sum_{x=-4}^{4} x = 0.$$

Let $Z = \big(X - \mathbf{E}[X]\big)^2 = X^2$. As in Example 2.1, we obtain

$$p_Z(z) = \begin{cases} 2/9 & \text{if } z = 1, 4, 9, 16, \\ 1/9 & \text{if } z = 0, \\ 0 & \text{otherwise.} \end{cases}$$

The variance of X is then obtained by

$$\mathrm{var}(X) = \mathbf{E}[Z] = \sum_z z p_Z(z) = 0 \cdot \frac{1}{9} + 1 \cdot \frac{2}{9} + 4 \cdot \frac{2}{9} + 9 \cdot \frac{2}{9} + 16 \cdot \frac{2}{9} = \frac{60}{9}.$$

It turns out that there is an easier method to calculate $\mathrm{var}(X)$, which uses the PMF of X but *does not require the PMF of* $\big(X - \mathbf{E}[X]\big)^2$. This method is based on the following rule.

Expected Value Rule for Functions of Random Variables

Let X be a random variable with PMF $p_X(x)$, and let $g(X)$ be a real-valued function of X. Then, the expected value of the random variable $g(X)$ is given by

$$\mathbf{E}\big[g(X)\big] = \sum_x g(x) p_X(x).$$

To verify this rule, we use the formula $p_Y(y) = \sum_{\{x \mid g(x)=y\}} p_X(x)$ derived in the preceding section, we have

$$\begin{aligned}
\mathbf{E}\big[g(X)\big] &= \mathbf{E}[Y] \\
&= \sum_y y p_Y(y) \\
&= \sum_y y \sum_{\{x \mid g(x)=y\}} p_X(x) \\
&= \sum_y \sum_{\{x \mid g(x)=y\}} y p_X(x) \\
&= \sum_y \sum_{\{x \mid g(x)=y\}} g(x) p_X(x) \\
&= \sum_x g(x) p_X(x).
\end{aligned}$$

Using the expected value rule, we can write the variance of X as

$$\mathrm{var}(X) = \mathbf{E}\left[\big(X - \mathbf{E}[X]\big)^2\right] = \sum_x \big(x - \mathbf{E}[X]\big)^2 p_X(x).$$

Similarly, the nth moment is given by

$$\mathbf{E}[X^n] = \sum_x x^n p_X(x),$$

and there is no need to calculate the PMF of X^n.

Example 2.3. (Continued) For the random variable X with PMF

$$p_X(x) = \begin{cases} 1/9 & \text{if } x \text{ is an integer in the range } [-4, 4], \\ 0 & \text{otherwise,} \end{cases}$$

we have

$$\begin{aligned}
\mathrm{var}(X) &= \mathbf{E}\left[\big(X - \mathbf{E}[X]\big)^2\right] \\
&= \sum_x \big(x - \mathbf{E}[X]\big)^2 p_X(x) \\
&= \frac{1}{9} \sum_{x=-4}^{4} x^2 \qquad \text{since } \mathbf{E}[X] = 0 \\
&= \frac{1}{9}(16 + 9 + 4 + 1 + 0 + 1 + 4 + 9 + 16) \\
&= \frac{60}{9},
\end{aligned}$$

which is consistent with the result obtained earlier.

As we have noted earlier, the variance is always nonnegative, but could it be zero? Since every term in the formula $\sum_x \big(x - \mathbf{E}[X]\big)^2 p_X(x)$ for the variance is nonnegative, the sum is zero if and only if $\big(x - \mathbf{E}[X]\big)^2 p_X(x) = 0$ for every x. This condition implies that for any x with $p_X(x) > 0$, we must have $x = \mathbf{E}[X]$ and the random variable X is not really "random": its experimental value is equal to the mean $\mathbf{E}[X]$, with probability 1.

Variance

The variance $\mathrm{var}(X)$ of a random variable X is defined by

$$\mathrm{var}(X) = \mathbf{E}\Big[\big(X - \mathbf{E}[X]\big)^2\Big]$$

and can be calculated as

$$\mathrm{var}(X) = \sum_x \big(x - \mathbf{E}[X]\big)^2 p_X(x).$$

It is always nonnegative. Its square root is denoted by σ_X and is called the **standard deviation**.

Let us now use the expected value rule for functions in order to derive some important properties of the mean and the variance. We start with a random variable X and define a new random variable Y, of the form

$$Y = aX + b,$$

where a and b are given scalars. Let us derive the mean and the variance of the linear function Y. We have

$$\mathbf{E}[Y] = \sum_x (ax + b)p_X(x) = a\sum_x xp_X(x) + b\sum_x p_X(x) = a\mathbf{E}[X] + b.$$

Furthermore,

$$\begin{aligned}
\mathrm{var}(Y) &= \sum_x \big(ax + b - \mathbf{E}[aX + b]\big)^2 p_X(x) \\
&= \sum_x \big(ax + b - a\mathbf{E}[X] - b\big)^2 p_X(x) \\
&= a^2 \sum_x \big(x - \mathbf{E}[X]\big)^2 p_X(x) \\
&= a^2 \mathrm{var}(X).
\end{aligned}$$

Mean and Variance of a Linear Function of a Random Variable

Let X be a random variable and let

$$Y = aX + b,$$

where a and b are given scalars. Then,

$$\mathbf{E}[Y] = a\mathbf{E}[X] + b, \qquad \mathrm{var}(Y) = a^2\mathrm{var}(X).$$

Let us also give a convenient formula for the variance of a random variable X with given PMF.

Variance in Terms of Moments Expression

$$\mathrm{var}(X) = \mathbf{E}[X^2] - \big(\mathbf{E}[X]\big)^2.$$

This expression is verified as follows:

$$\begin{aligned}
\mathrm{var}(X) &= \sum_x \big(x - \mathbf{E}[X]\big)^2 p_X(x) \\
&= \sum_x \Big(x^2 - 2x\mathbf{E}[X] + \big(\mathbf{E}[X]\big)^2\Big) p_X(x) \\
&= \sum_x x^2 p_X(x) - 2\mathbf{E}[X] \sum_x x p_X(x) + \big(\mathbf{E}[X]\big)^2 \sum_x p_X(x) \\
&= \mathbf{E}[X^2] - 2\big(\mathbf{E}[X]\big)^2 + \big(\mathbf{E}[X]\big)^2 \\
&= \mathbf{E}[X^2] - \big(\mathbf{E}[X]\big)^2.
\end{aligned}$$

We will now derive the mean and the variance of a few important random variables.

Example 2.4. Mean and Variance of the Bernoulli. Consider the experiment of tossing a biased coin, which comes up a head with probability p and a tail with probability $1 - p$, and the Bernoulli random variable X with PMF

$$p_X(k) = \begin{cases} p & \text{if } k = 1, \\ 1 - p & \text{if } k = 0. \end{cases}$$

Its mean, second moment, and variance are given by the following calculations:

$$\mathbf{E}[X] = 1 \cdot p + 0 \cdot (1-p) = p,$$
$$\mathbf{E}[X^2] = 1^2 \cdot p + 0 \cdot (1-p) = p,$$
$$\text{var}(X) = \mathbf{E}[X^2] - \big(\mathbf{E}[X]\big)^2 = p - p^2 = p(1-p).$$

Example 2.5. Discrete Uniform Random Variable. What is the mean and variance of the roll of a fair six-sided die? If we view the result of the roll as a random variable X, its PMF is

$$p_X(k) = \begin{cases} 1/6 & \text{if } k = 1,2,3,4,5,6, \\ 0 & \text{otherwise.} \end{cases}$$

Since the PMF is symmetric around 3.5, we conclude that $\mathbf{E}[X] = 3.5$. Regarding the variance, we have

$$\begin{aligned} \text{var}(X) &= \mathbf{E}[X^2] - \big(\mathbf{E}[X]\big)^2 \\ &= \frac{1}{6}(1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2) - (3.5)^2, \end{aligned}$$

which yields $\text{var}(X) = 35/12$.

The above random variable is a special case of a **discrete uniformly distributed** random variable (or **discrete uniform** for short), which by definition, takes one out of a range of contiguous integer values, with equal probability. More precisely, this random variable has a PMF of the form

$$p_X(k) = \begin{cases} \dfrac{1}{b-a+1} & \text{if } k = a, a+1, \ldots, b, \\ 0 & \text{otherwise,} \end{cases}$$

where a and b are two integers with $a < b$; see Fig. 2.9.

The mean is

$$\mathbf{E}[X] = \frac{a+b}{2},$$

as can be seen by inspection, since the PMF is symmetric around $(a+b)/2$. To calculate the variance of X, we first consider the simpler case where $a = 1$ and $b = n$. It can be verified by induction on n that

$$\mathbf{E}[X^2] = \frac{1}{n}\sum_{k=1}^{n} k^2 = \frac{1}{6}(n+1)(2n+1).$$

We leave the verification of this as an exercise for the reader. The variance can now be obtained in terms of the first and second moments

$$\begin{aligned} \text{var}(X) &= \mathbf{E}[X^2] - \big(\mathbf{E}[X]\big)^2 \\ &= \frac{1}{6}(n+1)(2n+1) - \frac{1}{4}(n+1)^2 \\ &= \frac{1}{12}(n+1)(4n+2-3n-3) \\ &= \frac{n^2-1}{12}. \end{aligned}$$

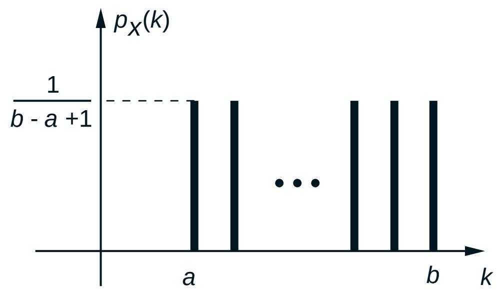

Figure 2.9: PMF of the discrete random variable that is uniformly distributed between two integers a and b. Its mean and variance are

$$\mathbf{E}[X] = \frac{a+b}{2}, \qquad \mathrm{var}(X) = \frac{(b-a)(b-a+2)}{12}.$$

For the case of general integers a and b, we note that the uniformly distributed random variable over $[a, b]$ has the same variance as the uniformly distributed random variable over the interval $[1, b - a + 1]$, since these two random variables differ by the constant $a - 1$. Therefore, the desired variance is given by the above formula with $n = b - a + 1$, which yields

$$\mathrm{var}(X) = \frac{(b-a+1)^2 - 1}{12} = \frac{(b-a)(b-a+2)}{12}.$$

Example 2.6. The Mean of the Poisson. The mean of the Poisson PMF

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

can be calculated is follows:

$$\begin{aligned} E[X] &= \sum_{k=0}^{\infty} k e^{-\lambda}\frac{\lambda^k}{k!} \\ &= \sum_{k=1}^{\infty} k e^{-\lambda}\frac{\lambda^k}{k!} \qquad \text{the } k = 0 \text{ term is zero} \\ &= \lambda \sum_{k=1}^{\infty} e^{-\lambda}\frac{\lambda^{k-1}}{(k-1)!} \\ &= \lambda \sum_{m=0}^{\infty} e^{-\lambda}\frac{\lambda^m}{m!} \qquad \text{let } m = k - 1 \\ &= \lambda. \end{aligned}$$

The last equality is obtained by noting that $\sum_{m=0}^{\infty} e^{-\lambda}\frac{\lambda^m}{m!} = \sum_{m=0}^{\infty} p_X(m) = 1$ is the normalization property for the Poisson PMF.

A similar calculation shows that the variance of a Poisson random variable is also λ (see the solved problems). We will have the occasion to derive this fact in a number of different ways in later chapters.

Expected values often provide a convenient vehicle for choosing optimally between several candidate decisions that result in different expected rewards. If we view the expected reward of a decision as its "average payoff over a large number of trials," it is reasonable to choose a decision with maximum expected reward. The following is an example.

Example 2.7. The Quiz Problem. This example, when generalized appropriately, is a prototypical model for optimal scheduling of a collection of tasks that have uncertain outcomes.

Consider a quiz game where a person is given two questions and must decide which question to answer first. Question 1 will be answered correctly with probability 0.8, and the person will then receive as prize \$100, while question 2 will be answered correctly with probability 0.5, and the person will then receive as prize \$200. If the first question attempted is answered incorrectly, the quiz terminates, i.e., the person is not allowed to attempt the second question. If the first question is answered correctly, the person is allowed to attempt the second question. Which question should be answered first to maximize the expected value of the total prize money received?

The answer is not obvious because there is a tradeoff: attempting first the more valuable but also more difficult question 2 carries the risk of never getting a chance to attempt the easier question 1. Let us view the total prize money received as a random variable X, and calculate the expected value $\mathbf{E}[X]$ under the two possible question orders (cf. Fig. 2.10):

Figure 2.10: Sequential description of the sample space of the quiz problem for the two cases where we answer question 1 or question 2 first.

(a) *Answer question 1 first*: Then the PMF of X is (cf. the left side of Fig. 2.10)

$$p_X(0) = 0.2, \qquad p_X(100) = 0.8 \cdot 0.5, \qquad p_X(300) = 0.8 \cdot 0.5,$$

and we have

$$\mathbf{E}[X] = 0.8 \cdot 0.5 \cdot 100 + 0.8 \cdot 0.5 \cdot 300 = \$160.$$

(b) *Answer question 2 first*: Then the PMF of X is (cf. the right side of Fig. 2.10)

$$p_X(0) = 0.5, \qquad p_X(200) = 0.5 \cdot 0.2, \qquad p_X(300) = 0.5 \cdot 0.8,$$

and we have

$$\mathbf{E}[X] = 0.5 \cdot 0.2 \cdot 200 + 0.5 \cdot 0.8 \cdot 300 = \$140.$$

Thus, it is preferable to attempt the easier question 1 first.

Let us now generalize the analysis. Denote by p_1 and p_2 the probabilities of correctly answering questions 1 and 2, respectively, and by v_1 and v_2 the corresponding prizes. If question 1 is answered first, we have

$$\mathbf{E}[X] = p_1(1 - p_2)v_1 + p_1p_2(v_1 + v_2) = p_1v_1 + p_1p_2v_2,$$

while if question 2 is answered first, we have

$$\mathbf{E}[X] = p_2(1 - p_1)v_2 + p_2p_1(v_2 + v_1) = p_2v_2 + p_2p_1v_1.$$

It is thus optimal to answer question 1 first if and only if

$$p_1v_1 + p_1p_2v_2 \geq p_2v_2 + p_2p_1v_1,$$

or equivalently, if

$$\frac{p_1v_1}{1 - p_1} \geq \frac{p_2v_2}{1 - p_2}.$$

Thus, it is optimal to order the questions in decreasing value of the expression $pv/(1 - p)$, which provides a convenient index of quality for a question with probability of correct answer p and value v. Interestingly, this rule generalizes to the case of more than two questions (see the end-of-chapter problems).

We finally illustrate by example a common pitfall: unless $g(X)$ is a linear function, it is not generally true that $\mathbf{E}\big[g(X)\big]$ is equal to $g\big(\mathbf{E}[X]\big)$.

Example 2.8. Average Speed Versus Average Time. If the weather is good (which happens with probability 0.6), Alice walks the 2 miles to class at a speed of $V = 5$ miles per hour, and otherwise drives her motorcycle at a speed of $V = 30$ miles per hour. What is the mean of the time T to get to class?

The correct way to solve the problem is to first derive the PMF of T,

$$p_T(t) = \begin{cases} 0.6 & \text{if } t = 2/5 \text{ hours}, \\ 0.4 & \text{if } t = 2/30 \text{ hours}, \end{cases}$$

and then calculate its mean by

$$\mathbf{E}[T] = 0.6 \cdot \frac{2}{5} + 0.4 \cdot \frac{2}{30} = \frac{4}{15} \text{ hours}.$$

However, it is wrong to calculate the mean of the speed V,

$$\mathbf{E}[V] = 0.6 \cdot 5 + 0.4 \cdot 30 = 15 \text{ miles per hour},$$

and then claim that the mean of the time T is

$$\frac{2}{\mathbf{E}[V]} = \frac{2}{15} \text{ hours}.$$

To summarize, in this example we have

$$T = \frac{2}{V}, \qquad \text{and} \quad \mathbf{E}[T] = \mathbf{E}\left[\frac{2}{V}\right] \neq \frac{2}{\mathbf{E}[V]}.$$

2.5 JOINT PMFS OF MULTIPLE RANDOM VARIABLES

Probabilistic models often involve several random variables of interest. For example, in a medical diagnosis context, the results of several tests may be significant, or in a networking context, the workloads of several gateways may be of interest. All of these random variables are associated with the same experiment, sample space, and probability law, and their values may relate in interesting ways. This motivates us to consider probabilities involving simultaneously the numerical values of several random variables and to investigate their mutual couplings. In this section, we will extend the concepts of PMF and expectation developed so far to multiple random variables. Later on, we will also develop notions of conditioning and independence that closely parallel the ideas discussed in Chapter 1.

Consider two discrete random variables X and Y associated with the same experiment. The **joint** PMF of X and Y is defined by

$$p_{X,Y}(x, y) = \mathbf{P}(X = x, Y = y)$$

for all pairs of numerical values (x, y) that X and Y can take. Here and elsewhere, we will use the abbreviated notation $\mathbf{P}(X = x, Y = y)$ instead of the more precise notations $\mathbf{P}(\{X = x\} \cap \{Y = y\})$ or $\mathbf{P}(X = x \text{ and } Y = x)$.

The joint PMF determines the probability of any event that can be specified in terms of the random variables X and Y. For example if A is the set of all pairs (x, y) that have a certain property, then

$$\mathbf{P}\big((X,Y) \in A\big) = \sum_{(x,y)\in A} p_{X,Y}(x,y).$$

In fact, we can calculate the PMFs of X and Y by using the formulas

$$p_X(x) = \sum_y p_{X,Y}(x,y), \qquad p_Y(y) = \sum_x p_{X,Y}(x,y).$$

The formula for $p_X(x)$ can be verified using the calculation

$$\begin{aligned} p_X(x) &= \mathbf{P}(X = x) \\ &= \sum_y \mathbf{P}(X = x, Y = y) \\ &= \sum_y p_{X,Y}(x,y), \end{aligned}$$

where the second equality follows by noting that the event $\{X = x\}$ is the union of the disjoint events $\{X = x, Y = y\}$ as y ranges over all the different values of Y. The formula for $p_Y(y)$ is verified similarly. We sometimes refer to p_X and p_Y as the **marginal** PMFs, to distinguish them from the joint PMF.

The example of Fig. 2.11 illustrates the calculation of the marginal PMFs from the joint PMF by using the **tabular method**. Here, the joint PMF of X and Y is arranged in a two-dimensional table, and **the marginal PMF of X or Y at a given value is obtained by adding the table entries along a corresponding column or row**, respectively.

Functions of Multiple Random Variables

When there are multiple random variables of interest, it is possible to generate new random variables by considering functions involving several of these random variables. In particular, a function $Z = g(X, Y)$ of the random variables X and Y defines another random variable. Its PMF can be calculated from the joint PMF $p_{X,Y}$ according to

$$p_Z(z) = \sum_{\{(x,y)\,|\,g(x,y)=z\}} p_{X,Y}(x,y).$$

Furthermore, the expected value rule for functions naturally extends and takes the form

$$\mathbf{E}\big[g(X,Y)\big] = \sum_{x,y} g(x,y) p_{X,Y}(x,y).$$

The verification of this is very similar to the earlier case of a function of a single random variable. In the special case where g is linear and of the form $aX+bY+c$, where a, b, and c are given scalars, we have

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

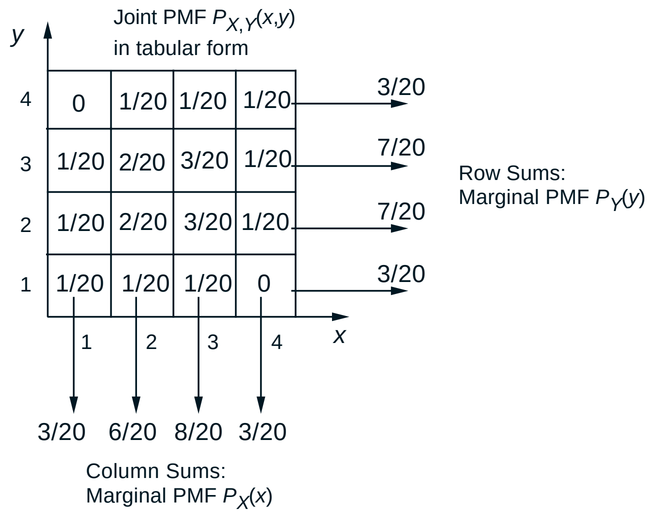

Figure 2.11: Illustration of the tabular method for calculating marginal PMFs from joint PMFs. The joint PMF is represented by a table, where the number in each square (x, y) gives the value of $p_{X,Y}(x, y)$. To calculate the marginal PMF $p_X(x)$ for a given value of x, we add the numbers in the column corresponding to x. For example $p_X(2) = 8/20$. Similarly, to calculate the marginal PMF $p_Y(y)$ for a given value of y, we add the numbers in the row corresponding to y. For example $p_Y(2) = 5/20$.

More than Two Random Variables

The joint PMF of three random variables X, Y, and Z is defined in analogy with the above as

$$p_{X,Y,Z}(x, y, z) = \mathbf{P}(X = x, Y = y, Z = z),$$

for all possible triplets of numerical values (x, y, z). Corresponding marginal PMFs are analogously obtained by equations such as

$$p_{X,Y}(x, y) = \sum_z p_{X,Y,Z}(x, y, z),$$

and

$$p_X(x) = \sum_y \sum_z p_{X,Y,Z}(x, y, z).$$

The expected value rule for functions takes the form

$$\mathbf{E}\big[g(X, Y, Z)\big] = \sum_{x,y,z} g(x, y, z) p_{X,Y,Z}(x, y, z),$$

and if g is linear and of the form $aX + bY + cZ + d$, then

$$\mathbf{E}[aX + bY + cZ + d] = a\mathbf{E}[X] + b\mathbf{E}[Y] + c\mathbf{E}[Z] + d.$$

Furthermore, there are obvious generalizations of the above to more than three random variables. For example, for any random variables $X_1, X_2, \ldots, X_n$ and any scalars $a_1, a_2, \ldots, a_n$, we have

$$\mathbf{E}[a_1X_1 + a_2X_2 + \cdots + a_nX_n] = a_1\mathbf{E}[X_1] + a_2\mathbf{E}[X_2] + \cdots + a_n\mathbf{E}[X_n].$$

Example 2.9. Mean of the Binomial. Your probability class has 300 students and each student has probability 1/3 of getting an A, independently of any other student. What is the mean of X, the number of students that get an A? Let

$$X_i = \begin{cases} 1 & \text{if the } i\text{th student gets an A,} \\ 0 & \text{otherwise.} \end{cases}$$

Thus $X_1, X_2, \ldots, X_n$ are Bernoulli random variables with common mean $p = 1/3$ and variance $p(1-p) = (1/3)(2/3) = 2/9$. Their sum

$$X = X_1 + X_2 + \cdots + X_n$$

is the number of students that get an A. Since X is the number of "successes" in n independent trials, it is a binomial random variable with parameters n and p.

Using the linearity of X as a function of the X_i, we have

$$\mathbf{E}[X] = \sum_{i=1}^{300} \mathbf{E}[X_i] = \sum_{i=1}^{300} \frac{1}{3} = 300 \cdot \frac{1}{3} = 100.$$

If we repeat this calculation for a general number of students n and probability of A equal to p, we obtain

$$\mathbf{E}[X] = \sum_{i=1}^{n} \mathbf{E}[X_i] = \sum_{i=1}^{n} p = np,$$

Example 2.10. The Hat Problem. Suppose that n people throw their hats in a box and then each picks up one hat at random. What is the expected value of X, the number of people that get back their own hat?

For the ith person, we introduce a random variable X_i that takes the value 1 if the person selects his/her own hat, and takes the value 0 otherwise. Since $\mathbf{P}(X_i = 1) = 1/n$ and $\mathbf{P}(X_i = 0) = 1 - 1/n$, the mean of X_i is

$$\mathbf{E}[X_i] = 1 \cdot \frac{1}{n} + 0 \cdot \left(1 - \frac{1}{n}\right) = \frac{1}{n}.$$

We now have

$$X = X_1 + X_2 + \cdots + X_n,$$

so that

$$\mathbf{E}[X] = \mathbf{E}[X_1] + \mathbf{E}[X_2] + \cdots + \mathbf{E}[X_n] = n \cdot \frac{1}{n} = 1.$$

Summary of Facts About Joint PMFs

Let X and Y be random variables associated with the same experiment.

- The joint PMF of X and Y is defined by

$$p_{X,Y}(x,y) = \mathbf{P}(X = x, Y = y).$$

- The marginal PMFs of X and Y can be obtained from the joint PMF, using the formulas

$$p_X(x) = \sum_y p_{X,Y}(x,y), \qquad p_Y(y) = \sum_x p_{X,Y}(x,y).$$

- A function $g(X,Y)$ of X and Y defines another random variable, and

$$\mathbf{E}\big[g(X,Y)\big] = \sum_{x,y} g(x,y)p_{X,Y}(x,y).$$

  If g is linear, of the form $aX + bY + c$, we have

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

- The above have natural extensions to the case where more than two random variables are involved.

2.6 CONDITIONING

If we have a probabilistic model and we are also told that a certain event A has occurred, we can capture this knowledge by employing the conditional instead of the original (unconditional) probabilities. As discussed in Chapter 1, conditional probabilities are like ordinary probabilities (satisfy the three axioms) except that they refer to a new universe in which event A is known to have occurred. In the same spirit, we can talk about conditional PMFs which provide the probabilities of the possible values of a random variable, conditioned on the occurrence of some event. This idea is developed in this section. In reality though, there is not much that is new, only an elaboration of concepts that are familiar from Chapter 1, together with a fair dose of new notation.

Conditioning a Random Variable on an Event

The **conditional PMF** of a random variable X, conditioned on a particular event A with $\mathbf{P}(A) > 0$, is defined by

$$p_{X|A}(x) = \mathbf{P}(X = x \mid A) = \frac{\mathbf{P}\big(\{X = x\} \cap A\big)}{\mathbf{P}(A)}.$$

Note that the events $\{X = x\} \cap A$ are disjoint for different values of x, their union is A, and, therefore,

$$\mathbf{P}(A) = \sum_x \mathbf{P}\big(\{X = x\} \cap A\big).$$

Combining the above two formulas, we see that

$$\sum_x p_{X|A}(x) = 1,$$

so $p_{X|A}$ is a legitimate PMF.

As an example, let X be the roll of a die and let A be the event that the roll is an even number. Then, by applying the preceding formula, we obtain

$$\begin{aligned} p_{X|A}(x) &= \mathbf{P}(X = x \mid \text{roll is even}) \\ &= \frac{\mathbf{P}(X = x \text{ and } X \text{ is even})}{\mathbf{P}(\text{roll is even})} \\ &= \begin{cases} 1/3 & \text{if } x = 2, 4, 6, \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

The conditional PMF is calculated similar to its unconditional counterpart: to obtain $p_{X|A}(x)$, we add the probabilities of the outcomes that give rise to $X = x$ **and** belong to the conditioning event A, and then normalize by dividing with $\mathbf{P}(A)$ (see Fig. 2.12).

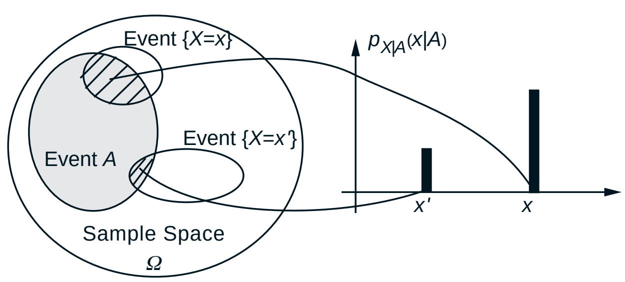

Figure 2.12: Visualization and calculation of the conditional PMF $p_{X|A}(x)$. For each x, we add the probabilities of the outcomes in the intersection $\{X = x\} \cap A$ and normalize by diving with $\mathbf{P}(A)$.

Conditioning one Random Variable on Another

Let X and Y be two random variables associated with the same experiment. If we know that the experimental value of Y is some particular y (with $p_Y(y) > 0$), this provides partial knowledge about the value of X. This knowledge is captured by the **conditional PMF** $p_{X|Y}$ of X given Y, which is defined by specializing the definition of $p_{X|A}$ to events A of the form $\{Y = y\}$:

$$p_{X|Y}(x \mid y) = \mathbf{P}(X = x \mid Y = y).$$

Using the definition of conditional probabilities, we have

$$p_{X|Y}(x \mid y) = \frac{\mathbf{P}(X = x, Y = y)}{\mathbf{P}(Y = y)} = \frac{p_{X,Y}(x, y)}{p_Y(y)}.$$

Let us fix some y, with $p_Y(y) > 0$ and consider $p_{X|Y}(x \mid y)$ as a function of x. This function is a valid PMF for X: it assigns nonnegative values to each possible x, and these values add to 1. Furthermore, this function of x, has the same shape as $p_{X,Y}(x, y)$ except that it is normalized by dividing with $p_Y(y)$, which enforces the normalization property

$$\sum_x p_{X|Y}(x \mid y) = 1.$$

Figure 2.13 provides a visualization of the conditional PMF.

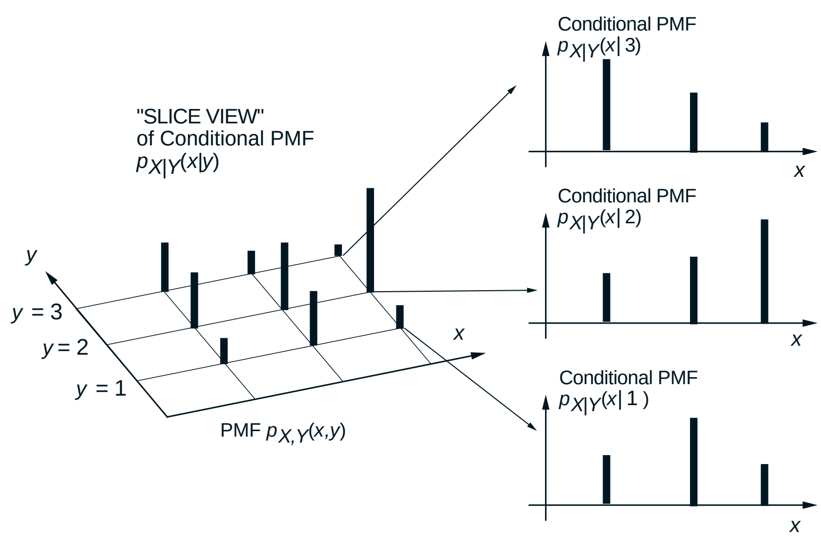

Figure 2.13: Visualization of the conditional PMF $p_{X|Y}(x \mid y)$. For each y, we view the joint PMF along the slice $Y = y$ and renormalize so that

$$\sum_x p_{X|Y}(x \mid y) = 1.$$

The conditional PMF is often convenient for the calculation of the joint PMF, using a sequential approach and the formula

$$p_{X,Y}(x,y) = p_Y(y)p_{X|Y}(x \,|\, y),$$

or its counterpart

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

This method is entirely similar to the use of the multiplication rule from Chapter 1. The following examples provide an illustration.

Example 2.11. Professor May B. Right often has her facts wrong, and answers each of her students' questions incorrectly with probability 1/4, independently of other questions. In each lecture May is asked 0, 1, or 2 questions with equal probability 1/3. Let X and Y be the number of questions May is asked and the number of questions she answers wrong in a given lecture, respectively. To construct the joint PMF $p_{X,Y}(x,y)$, we need to calculate all the probabilities $\mathbf{P}(X = x, Y = y)$ for all combinations of values of x and y. This can be done by using a sequential description of the experiment and the multiplication rule $p_{X,Y}(x,y) = p_Y(y)p_{X|Y}(x \,|\, y)$, as shown in Fig. 2.14. For example, for the case where one question is asked and is answered wrong, we have

$$p_{X,Y}(1,1) = p_X(x)p_{Y|X}(y \,|\, x) = \frac{1}{4} \cdot \frac{1}{3} = \frac{1}{12}.$$

The joint PMF can be represented by a two-dimensional table, as shown in Fig. 2.14. It can be used to calculate the probability of any event of interest. For instance, we have

$$\begin{aligned}\mathbf{P}(\text{at least one wrong answer}) &= p_{X,Y}(1,1) + p_{X,Y}(2,1) + p_{X,Y}(2,2)\\ &= \frac{4}{48} + \frac{6}{48} + \frac{1}{48}.\end{aligned}$$

.

Example 2.12. Consider four independent rolls of a 6-sided die. Let X be the number of 1's and let Y be the number of 2's obtained. What is the joint PMF of X and Y?

The marginal PMF p_Y is given by the binomial formula

$$p_Y(y) = \binom{4}{y}\left(\frac{1}{6}\right)^y\left(\frac{5}{6}\right)^{4-y}, \qquad y = 0, 1, \ldots, 4.$$

To compute the conditional PMF $p_{X|Y}$, note that given that $Y = y$, X is the number of 1's in the remaining $4 - y$ rolls, each of which can take the 5 values

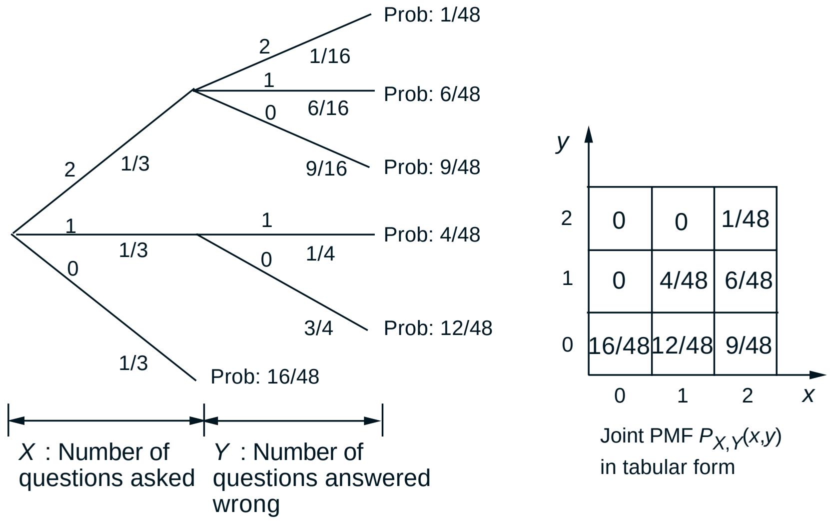

Figure 2.14: Calculation of the joint PMF $p_{X,Y}(x,y)$ in Example 2.11.

$1, 3, 4, 5, 6$ with equal probability $1/5$. Thus, the conditional PMF $p_{X|Y}$ is binomial with parameters $4-y$ and $p=1/5$:

$$p_{X|Y}(x\,|\,y) = \binom{4-y}{x}\left(\frac{1}{5}\right)^x\left(\frac{4}{5}\right)^{4-y-x},$$

for all x and y such that $x, y = 0, 1, \ldots, 4$, and $0 \le x+y \le 4$. The joint PMF is now given by

$$\begin{aligned} p_{X,Y}(x,y) &= p_Y(y)p_{X|Y}(x\,|\,y) \\ &= \binom{4}{y}\left(\frac{1}{6}\right)^y\left(\frac{5}{6}\right)^{4-y}\binom{4-y}{x}\left(\frac{1}{5}\right)^x\left(\frac{4}{5}\right)^{4-y-x}, \end{aligned}$$

for all nonnegative integers x and y such that $0 \le x+y \le 4$. For other values of x and y, we have $p_{X,Y}(x,y) = 0$.

The conditional PMF can also be used to calculate the marginal PMFs. In particular, we have by using the definitions,

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

This formula provides a divide-and-conquer method for calculating marginal PMFs. It is in essence identical to the total probability theorem given in Chapter 1, but cast in different notation. The following example provides an illustration.

Example 2.13. Consider a transmitter that is sending messages over a computer network. Let us define the following two random variables:

X : the travel time of a given message, Y : the length of the given message.

We know the PMF of the travel time of a message that has a given length, and we know the PMF of the message length. We want to find the (unconditional) PMF of the travel time of a message.

We assume that the length of a message can take two possible values: $y = 10^2$ bytes with probability 5/6, and $y = 10^4$ bytes with probability 1/6, so that

$$p_Y(y) = \begin{cases} 5/6 & \text{if } y = 10^2, \\ 1/6 & \text{if } y = 10^4. \end{cases}$$

We assume that the travel time X of the message depends on its length Y and the congestion level of the network at the time of transmission. In particular, the travel time is $10^{-4}Y$ secs with probability 1/2, $10^{-3}Y$ secs with probability 1/3, and $10^{-2}Y$ secs with probability 1/6. Thus, we have

$$p_{X|Y}(x \,|\, 10^2) = \begin{cases} 1/2 & \text{if } x = 10^{-2}, \\ 1/3 & \text{if } x = 10^{-1}, \\ 1/6 & \text{if } x = 1, \end{cases} \qquad p_{X|Y}(x \,|\, 10^4) = \begin{cases} 1/2 & \text{if } x = 1, \\ 1/3 & \text{if } x = 10, \\ 1/6 & \text{if } x = 100. \end{cases}$$

To find the PMF of X, we use the total probability formula

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

We obtain

$$p_X(10^{-2}) = \frac{5}{6} \cdot \frac{1}{2}, \qquad p_X(10^{-1}) = \frac{5}{6} \cdot \frac{1}{3}, \qquad p_X(1) = \frac{5}{6} \cdot \frac{1}{6} + \frac{1}{6} \cdot \frac{1}{2},$$

$$p_X(10) = \frac{1}{6} \cdot \frac{1}{3}, \qquad p_X(100) = \frac{1}{6} \cdot \frac{1}{6}.$$

Note finally that one can define conditional PMFs involving more than two random variables, as in $p_{X,Y|Z}(x, y \,|\, z)$ or $p_{X|Y,Z}(x \,|\, y, z)$. The concepts and methods described above generalize easily (see the end-of-chapter problems).

Summary of Facts About Conditional PMFs

Let X and Y be random variables associated with the same experiment.

- Conditional PMFs are similar to ordinary PMFs, but refer to a universe where the conditioning event is known to have occurred.
- The conditional PMF of X given an event A with $\mathbf{P}(A) > 0$, is defined by
$$p_{X|A}(x) = \mathbf{P}(X = x \mid A)$$
and satisfies
$$\sum_x p_{X|A}(x) = 1.$$
- The conditional PMF of X given $Y = y$ is related to the joint PMF by
$$p_{X,Y}(x, y) = p_Y(y) p_{X|Y}(x \mid y).$$
This is analogous to the multiplication rule for calculating probabilities and can be used to calculate the joint PMF from the conditional PMF.
- The conditional PMF of X given Y can be used to calculate the marginal PMFs with the formula
$$p_X(x) = \sum_y p_Y(y) p_{X|Y}(x \mid y).$$
This is analogous to the divide-and-conquer approach for calculating probabilities using the total probability theorem.
- There are natural extensions to the above involving more than two random variables.

Conditional Expectation

A conditional PMF can be thought of as an ordinary PMF over a new universe determined by the conditioning event. In the same spirit, a conditional expectation is the same as an ordinary expectation, except that it refers to the new universe, and all probabilities and PMFs are replaced by their conditional counterparts. We list the main definitions and relevant facts below.

Summary of Facts About Conditional Expectations

Let X and Y be random variables associated with the same experiment.

- The conditional expectation of X given an event A with $\mathbf{P}(A) > 0$, is defined by
$$\mathbf{E}[X \mid A] = \sum_x x p_{X|A}(x \mid A).$$
For a function $g(X)$, it is given by
$$\mathbf{E}\big[g(X) \mid A\big] = \sum_x g(x) p_{X|A}(x \mid A).$$
- The conditional expectation of X given a value y of Y is defined by
$$\mathbf{E}[X \mid Y = y] = \sum_x x p_{X|Y}(x \mid y).$$
- We have
$$\mathbf{E}[X] = \sum_y p_Y(y) \mathbf{E}[X \mid Y = y].$$
This is the **total expectation theorem**.
- Let $A_1, \ldots, A_n$ be disjoint events that form a partition of the sample space, and assume that $\mathbf{P}(A_i) > 0$ for all i. Then,
$$\mathbf{E}[X] = \sum_{i=1}^{n} \mathbf{P}(A_i) \mathbf{E}[X \mid A_i].$$

Let us verify the total expectation theorem, which basically says that "the unconditional average can be obtained by averaging the conditional averages." The theorem is derived using the total probability formula

$$p_X(x) = \sum_y p_Y(y) p_{X|Y}(x \mid y)$$

and the calculation

$$\begin{aligned}\mathbf{E}[X] &= \sum_x x p_X(x)\\ &= \sum_x x \sum_y p_Y(y) p_{X|Y}(x\,|\,y)\\ &= \sum_y p_Y(y) \sum_x x p_{X|Y}(x\,|\,y)\\ &= \sum_y p_Y(y)\mathbf{E}[X\,|\,Y=y].\end{aligned}$$

The relation $\mathbf{E}[X] = \sum_{i=1}^n \mathbf{P}(A_i)\mathbf{E}[X\,|\,A_i]$ can be verified by viewing it as a special case of the total expectation theorem. Let us introduce the random variable Y that takes the value i if and only if the event A_i occurs. Its PMF is given by

$$p_Y(i) = \begin{cases} \mathbf{P}(A_i) & \text{if } i = 1, 2, \ldots, n,\\ 0 & \text{otherwise.}\end{cases}$$

The total expectation theorem yields

$$\mathbf{E}[X] = \sum_{i=1}^n \mathbf{P}(A_i)\mathbf{E}[X\,|\,Y=i],$$

and since the event $\{Y = i\}$ is just A_i, we obtain the desired expression

$$\mathbf{E}[X] = \sum_{i=1}^n \mathbf{P}(A_i)\mathbf{E}[X\,|\,A_i].$$

The total expectation theorem is analogous to the total probability theorem. It can be used to calculate the unconditional expectation $\mathbf{E}[X]$ from the conditional PMF or expectation, using a divide-and-conquer approach.

Example 2.14. Messages transmitted by a computer in Boston through a data network are destined for New York with probability 0.5, for Chicago with probability 0.3, and for San Francisco with probability 0.2. The transit time X of a message is random. Its mean is 0.05 secs if it is destined for New York, 0.1 secs if it is destined for Chicago, and 0.3 secs if it is destined for San Francisco. Then, $\mathbf{E}[X]$ is easily calculated using the total expectation theorem as

$$\mathbf{E}[X] = 0.5 \cdot 0.05 + 0.3 \cdot 0.1 + 0.2 \cdot 0.3 = 0.115 \text{ secs.}$$

Example 2.15. Mean and Variance of the Geometric Random Variable. You write a software program over and over, and each time there is probability p

that it works correctly, independently from previous attempts. What is the mean and variance of X, the number of tries until the program works correctly?

We recognize X as a geometric random variable with PMF

$$p_X(k) = (1-p)^{k-1}p, \qquad k = 1, 2, \ldots.$$

The mean and variance of X are given by

$$\mathbf{E}[X] = \sum_{k=1}^{\infty} k(1-p)^{k-1}p, \qquad \mathrm{var}(X) = \sum_{k=1}^{\infty} (k - \mathbf{E}[X])^2 (1-p)^{k-1}p,$$

but evaluating these infinite sums is somewhat tedious. As an alternative, we will apply the total expectation theorem, with $A_1 = \{X = 1\} = \{\text{first try is a success}\}$, $A_2 = \{X > 1\} = \{\text{first try is a failure}\}$, and end up with a much simpler calculation.

If the first try is successful, we have $X = 1$, and

$$\mathbf{E}[X \mid X = 1] = 1.$$

If the first try fails ($X > 1$), we have wasted one try, and we are back where we started. So, the expected number of remaining tries is $\mathbf{E}[X]$, and

$$\mathbf{E}[X \mid X > 1] = 1 + \mathbf{E}[X].$$

Thus,

$$\begin{aligned}\mathbf{E}[X] &= \mathbf{P}(X = 1)\mathbf{E}[X \mid X = 1] + \mathbf{P}(X > 1)\mathbf{E}[X \mid X > 1] \\ &= p + (1-p)\big(1 + \mathbf{E}[X]\big),\end{aligned}$$

from which we obtain

$$\mathbf{E}[X] = \frac{1}{p}.$$

With similar reasoning, we also have

$$\mathbf{E}[X^2 \mid X = 1] = 1, \qquad \mathbf{E}[X^2 \mid X > 1] = \mathbf{E}\big[(1 + X)^2\big] = 1 + 2\mathbf{E}[X] + \mathbf{E}[X^2],$$

so that

$$\mathbf{E}[X^2] = p \cdot 1 + (1-p)\big(1 + 2\mathbf{E}[X] + \mathbf{E}[X^2]\big),$$

from which we obtain

$$\mathbf{E}[X^2] = \frac{1 + 2(1-p)\mathbf{E}[X]}{p},$$

and

$$\mathbf{E}[X^2] = \frac{2}{p^2} - \frac{1}{p}.$$

We conclude that

$$\mathrm{var}(X) = \mathbf{E}[X^2] - \big(\mathbf{E}[X]\big)^2 = \frac{2}{p^2} - \frac{1}{p} - \frac{1}{p^2} = \frac{1-p}{p^2}.$$

2.7 INDEPENDENCE

We now discuss concepts of independence related to random variables. These concepts are analogous to the concepts of independence between events (cf. Chapter 1). They are developed by simply introducing suitable events involving the possible values of various random variables, and by considering their independence.

Independence of a Random Variable from an Event

The independence of a random variable from an event is similar to the independence of two events. The idea is that knowing the occurrence of the conditioning event tells us nothing about the value of the random variable. More formally, we say that the random variable X is **independent of the event** A if

$$\mathbf{P}(X = x \text{ and } A) = \mathbf{P}(X = x)\mathbf{P}(A) = p_X(x)\mathbf{P}(A), \qquad \text{for all } x,$$

which is the same as requiring that the two events $\{X = x\}$ and A be independent, for any choice x. As long as $\mathbf{P}(A) > 0$, and using the definition $p_{X|A}(x) = \mathbf{P}(X = x \text{ and } A)/\mathbf{P}(A)$ of the conditional PMF, we see that independence is the same as the condition

$$p_{X|A}(x) = p_X(x), \qquad \text{for all } x.$$

Example 2.16. Consider two independent tosses of a fair coin. Let X be the number of heads and let A be the event that the number of heads is even. The (unconditional) PMF of X is

$$p_X(x) = \begin{cases} 1/4 & \text{if } x = 0, \\ 1/2 & \text{if } x = 1, \\ 1/4 & \text{if } x = 2, \end{cases}$$

and $\mathbf{P}(A) = 1/2$. The conditional PMF is obtained from the definition $p_{X|A}(x) = \mathbf{P}\big(X = x \text{ and } A\big)/\mathbf{P}(A)$:

$$p_{X|A}(x) = \begin{cases} 1/2 & \text{if } x = 0, \\ 0 & \text{if } x = 1, \\ 1/2 & \text{if } x = 2. \end{cases}$$

Clearly, X and A are not independent, since the PMFs p_X and $p_{X|A}$ are different. For an example of a random variable that is independent of A, consider the random variable that takes the value 0 if the first toss is a head, and the value 1 if the first toss is a tail. This is intuitively clear and can also be verified by using the definition of independence.

Independence of Random Variables

The notion of independence of two random variables is similar. We say that two **random variables** X and Y are **independent** if

$$p_{X,Y}(x,y) = p_X(x)\,p_Y(y), \qquad \text{for all } x, y.$$

This is the same as requiring that the two events $\{X = x\}$ and $\{Y = y\}$ be independent for every x and y. Finally, the formula $p_{X,Y}(x,y) = p_{X|Y}(x\,|\,y)p_Y(y)$ shows that independence is equivalent to the condition

$$p_{X|Y}(x\,|\,y) = p_X(x), \qquad \text{for all } y \text{ with } p_Y(y) > 0 \text{ and all } x.$$

Intuitively, independence means that the experimental value of Y tells us nothing about the value of X.

There is a similar notion of conditional independence of two random variables, given an event A with $\mathbf{P}(A > 0$. The conditioning event A defines a new universe and all probabilities (or PMFs) have to be replaced by their conditional counterparts. For example, X and Y are said to be **conditionally independent**, given a positive probability event A, if

$$\mathbf{P}(X = x, Y = y\,|\,A) = \mathbf{P}(X = x\,|\,A)\mathbf{P}(Y = y\,|\,A), \qquad \text{for all } x \text{ and } y,$$

or, in this chapter's notation,

$$p_{X,Y|A}(x,y) = p_{X|A}(x)p_{Y|A}(y), \qquad \text{for all } x \text{ and } y.$$

Once more, this is equivalent to

$$p_{X|Y,A}(x\,|\,y) = p_{X|A}(x) \qquad \text{for all } x \text{ and } y \text{ such that } p_{Y|A}(y) > 0.$$

As in the case of events (Section 1.4), conditional independence may not imply unconditional independence and vice versa. This is illustrated by the example in Fig. 2.15.

If X and Y are independent random variables, then

$$\mathbf{E}[XY] = \mathbf{E}[X]\,\mathbf{E}[Y],$$

as shown by the following calculation:

$$\begin{aligned}
\mathbf{E}[XY] &= \sum_x \sum_y xyp_{X,Y}(x,y) \\
&= \sum_x \sum_y xyp_X(x)p_Y(y) \qquad \text{by independence} \\
&= \sum_x xp_X(x) \sum_y yp_Y(y) \\
&= \mathbf{E}[X]\,\mathbf{E}[Y].
\end{aligned}$$

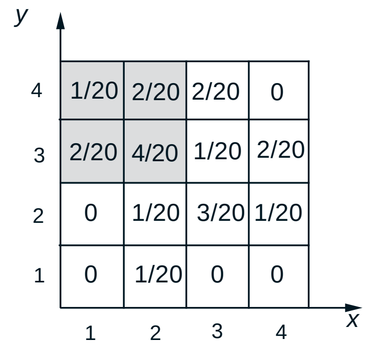

Figure 2.15: Example illustrating that conditional independence may not imply unconditional independence. For the PMF shown, the random variables X and Y are not independent. For example, we have

$$p_{X|Y}(1\,|\,1) = \mathbf{P}(X = 1\,|\,Y = 1) = 0 \neq \mathbf{P}(X = 1) = p_X(1).$$

On the other hand, conditional on the event $A = \{X \leq 2,\, Y \geq 3\}$ (the shaded set in the figure), the random variables X and Y can be seen to be independent. In particular, we have

$$p_{X|Y,A}(x\,|\,y) = \begin{cases} 1/3 & \text{if } x = 1, \\ 2/3 & \text{if } x = 2, \end{cases}$$

for both values $y = 3$ and $y = 4$.

A very similar calculation also shows that if X and Y are independent, then

$$\mathbf{E}\big[g(X)h(Y)\big] = \mathbf{E}\big[g(X)\big]\mathbf{E}\big[h(Y)\big],$$

for any functions g and h. In fact, this follows immediately once we realize that if X and Y are independent, then the same is true for $g(X)$ and $h(Y)$. This is intuitively clear and its formal verification is left as an end-of-chapter problem.

Consider now the sum $Z = X + Y$ of two independent random variables X and Y, and let us calculate the variance of Z. We have, using the relation $\mathbf{E}[X + Y] = \mathbf{E}[X] + \mathbf{E}[Y]$,

$$\begin{aligned}
\text{var}(Z) &= \mathbf{E}\Big[\big(X + Y - \mathbf{E}[X + Y]\big)^2\Big] \\
&= \mathbf{E}\Big[\big(X + Y - \mathbf{E}[X] - \mathbf{E}[Y]\big)^2\Big] \\
&= \mathbf{E}\Big[\Big(\big(X - \mathbf{E}[X]\big) + \big(Y - \mathbf{E}[Y]\big)\Big)^2\Big] \\
&= \mathbf{E}\Big[\big(X - \mathbf{E}[X]\big)^2\Big] + \mathbf{E}\Big[\big(Y - \mathbf{E}[Y]\big)^2\Big] \\
&\quad + 2\mathbf{E}\Big[\big(X - \mathbf{E}[X]\big)\big(Y - \mathbf{E}[Y]\big)\Big] \\
&= \mathbf{E}\Big[\big(X - \mathbf{E}[X]\big)^2\Big] + \mathbf{E}\Big[\big(Y - \mathbf{E}[Y]\big)^2\Big].
\end{aligned}$$

To justify the last equality, note that the random variables $X-\mathbf{E}[X]$ and $Y-\mathbf{E}[Y]$ are independent (they are functions of the independent random variables X and Y, respectively) and

$$\mathbf{E}\big[\big(X-\mathbf{E}[X]\big)\big(Y-\mathbf{E}[Y]\big)\big] = \mathbf{E}\big[\big(X-\mathbf{E}[X]\big)\big]\,\mathbf{E}\big[\big(Y-\mathbf{E}[Y]\big)\big] = 0.$$

We conclude that

$$\text{var}(Z) = \text{var}(X) + \text{var}(Y).$$

Thus, the variance of the sum of two **independent** random variables is equal to the sum of their variances. As an interesting contrast, note that the mean of the sum of two random variables is always equal to the sum of their means, even if they are not independent.

Summary of Facts About Independent Random Variables

Let A be an event, with $\mathbf{P}(A) > 0$, and let X and Y be random variables associated with the same experiment.

- X is independent of the event A if

  $$p_{X|A}(x) = p_X(x), \qquad \text{for all } x,$$

  that is, if for all x, the events $\{X = x\}$ and A are independent.

- X and Y are independent if for all possible pairs (x, y), the events $\{X = x\}$ and $\{Y = y\}$ are independent, or equivalently

  $$p_{X,Y}(x, y) = p_X(x)p_Y(y), \qquad \text{for all } x, y.$$

- If X and Y are independent random variables, then

  $$\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y].$$

  Furthermore, for any functions f and g, the random variables $g(X)$ and $h(Y)$ are independent, and we have

  $$\mathbf{E}\big[g(X)h(Y)\big] = \mathbf{E}\big[g(X)\big]\mathbf{E}\big[h(Y)\big].$$

- If X and Y are independent, then

  $$\text{var}[X + Y] = \text{var}(X) + \text{var}(Y).$$

Independence of Several Random Variables

All of the above have natural extensions to the case of more than two random variables. For example, three random variables X, Y, and Z are said to be independent if

$$p_{X,Y,Z}(x, y, z) = p_X(x)p_Y(y)p_Z(z), \qquad \text{for all } x, y, z.$$

If X, Y, and Z are independent random variables, then any three random variables of the form $f(X)$, $g(Y)$, and $h(Z)$, are also independent. Similarly, any two random variables of the form $g(X, Y)$ and $h(Z)$ are independent. On the other hand, two random variables of the form $g(X, Y)$ and $h(Y, Z)$ are usually not independent, because they are both affected by Y. Properties such as the above are intuitively clear if we interpret independence in terms of noninteracting (sub)experiments. They can be formally verified (see the end-of-chapter problems), but this is sometimes tedious. Fortunately, there is general agreement between intuition and what is mathematically correct. This is basically a testament that the definitions of independence we have been using adequately reflect the intended interpretation.

Another property that extends to multiple random variables is the following. If $X_1, X_2, \ldots, X_n$ are independent random variables, then

$$\text{var}(X_1 + X_2 + \cdots + X_n) = \text{var}(X_1) + \text{var}(X_2) + \cdots + \text{var}(X_n).$$

This can be verified by a calculation similar to the one for the case of two random variables and is left as an exercise for the reader.

Example 2.17. Variance of the Binomial. We consider n independent coin tosses, with each toss having probability p of coming up a head. For each i, we let X_i be the Bernoulli random variable which is equal to 1 if the ith toss comes up a head, and is 0 otherwise. Then, $X = X_1 + X_2 + \cdots + X_n$ is a binomial random variable. By the independence of the coin tosses, the random variables $X_1, \ldots, X_n$ are independent, and

$$\text{var}(X) = \sum_{i=1}^{n} \text{var}(X_i) = np(1 - p).$$

The formulas for the mean and variance of a weighted sum of random variables form the basis for many statistical procedures that estimate the mean of a random variable by averaging many independent samples. A typical case is illustrated in the following example.

Example 2.18. Mean and Variance of the Sample Mean. We wish to estimate the approval rating of a president, to be called C. To this end, we ask n

persons drawn at random from the voter population, and we let X_i be a random variable that encodes the response of the ith person:

$$X_i = \begin{cases} 1 & \text{if the } i\text{th person approves C's performance,} \\ 0 & \text{if the } i\text{th person disapproves C's performance.} \end{cases}$$

We model $X_1, X_2, \ldots, X_n$ as independent Bernoulli random variables with common mean p and variance $p(1-p)$. Naturally, we view p as the true approval rating of C. We "average" the responses and compute the **sample mean** S_n, defined as

$$S_n = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

Thus, S_n is the approval rating of C within our n-person sample.

We have, using the linearity of S_n as a function of the X_i,

$$\mathbf{E}[S_n] = \sum_{i=1}^{n} \frac{1}{n}\mathbf{E}[X_i] = \frac{1}{n}\sum_{i=1}^{n} p = p,$$

and making use of the independence of $X_1, \ldots, X_n$,

$$\text{var}(S_n) = \sum_{i=1}^{n} \frac{1}{n^2}\text{var}(X_i) = \frac{p(1-p)}{n}.$$

The sample mean S_n can be viewed as a "good" estimate of the approval rating. This is because it has the correct expected value, which is the approval rating p, and its accuracy, as reflected by its variance, improves as the sample size n increases.

Note that even if the random variables X_i are not Bernoulli, the same calculation yields

$$\text{var}(S_n) = \frac{\text{var}(X)}{n},$$

as long as the X_i are independent, with common mean $\mathbf{E}[X]$ and variance $\text{var}(X)$. Thus, again, the sample mean becomes a very good estimate (in terms of variance) of the true mean $\mathbf{E}[X]$, as the sample size n increases. We will revisit the properties of the sample mean and discuss them in much greater detail in Chapter 7, when we discuss the laws of large numbers.

Example 2.19. Estimating Probabilities by Simulation. In many practical situations, the analytical calculation of the probability of some event of interest is very difficult. However, if we have a physical or computer model that can generate outcomes of a given experiment in accordance with their true probabilities, we can use simulation to calculate with high accuracy the probability of any given event A. In particular, we independently generate with our model n outcomes, we record the number m that belong to the event A of interest, and we approximate $\mathbf{P}(A)$ by m/n. For example, to calculate the probability $p = \mathbf{P}(\text{Heads})$ of a biased coin, we flip the coin n times, and we approximate p with the ratio (number of heads recorded)$/n$.

To see how accurate this process is, consider n independent Bernoulli random variables $X_1, \ldots, X_n$, each with PMF

$$p_{X_i}(x_i) = \begin{cases} \mathbf{P}(A) & \text{if } x_i = 1, \\ 0 & \text{if } x_i = 0. \end{cases}$$

In a simulation context, X_i corresponds to the ith outcome, and takes the value 1 if the ith outcome belongs to the event A. The value of the random variable

$$X = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

is the estimate of $\mathbf{P}(A)$ provided by the simulation. According to Example 2.17, X has mean $\mathbf{P}(A)$ and variance $\mathbf{P}(A)\big(1 - \mathbf{P}(A)\big)/n$, so that for large n, it provides an accurate estimate of $\mathbf{P}(A)$.

2.8 SUMMARY AND DISCUSSION

Random variables provide the natural tools for dealing with probabilistic models in which the outcome determines certain numerical values of interest. In this chapter, we focused on discrete random variables, and developed the main concepts and some relevant tools. We also discussed several special random variables, and derived their PMF, mean, and variance, as summarized in the table that follows.

Summary of Results for Special Random Variables

Discrete Uniform over $[a, b]$**:**

$$p_X(k) = \begin{cases} \dfrac{1}{b - a + 1} & \text{if } k = a, a+1, \ldots, b, \\ 0 & \text{otherwise,} \end{cases}$$

$$\mathbf{E}[X] = \frac{a+b}{2}, \qquad \text{var}(X) = \frac{(b-a)(b-a+2)}{12}.$$

Bernoulli with Parameter p**:** (Describes the success or failure in a single trial.)

$$p_X(k) = \begin{cases} p & \text{if } k = 1, \\ 1 - p & \text{if } k = 0, \end{cases}$$

$$\mathbf{E}[X] = p, \qquad \text{var}(X) = p(1 - p).$$

Binomial with Parameters p and n: (Describes the number of successes in n independent Bernoulli trials.)

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, 1, \ldots, n,$$

$$\mathbf{E}[X] = np, \qquad \text{var}(X) = np(1-p).$$

Geometric with Parameter p: (Describes the number of trials until the first success, in a sequence of independent Bernoulli trials.)

$$p_X(k) = (1-p)^{k-1}p, \qquad k = 1, 2, \ldots,$$

$$\mathbf{E}[X] = \frac{1}{p}, \qquad \text{var}(X) = \frac{1-p}{p^2}.$$

Poisson with Parameter λ: (Approximates the binomial PMF when n is large, p is small, and $\lambda = np$.)

$$p_X(k) = e^{-\lambda}\frac{\lambda^k}{k!}, \qquad k = 0, 1, \ldots,$$

$$\mathbf{E}[X] = \lambda, \qquad \text{var}(X) = \lambda.$$

We also considered multiple random variables, and introduced their joint and conditional PMFs, and associated expected values. Conditional PMFs are often the starting point in probabilistic models and can be used to calculate other quantities of interest, such as marginal or joint PMFs and expectations, through a sequential or a divide-and-conquer approach. In particular, given the conditional PMF $p_{X|Y}(x \,|\, y)$:

(a) The joint PMF can be calculated by

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

This can be extended to the case of three or more random variables, as in

$$p_{X,Y,Z}(x, y, z) = p_Y(y)p_{Y|Z}(y \,|\, z)p_{X|Y,Z}(x \,|\, y, z),$$

and is analogous to the sequential tree-based calculation method using the multiplication rule, discussed in Chapter 1.

(b) The marginal PMF can be calculated by

$$p_X(x) = \sum_y p_Y(y)p_{X|Y}(x \,|\, y),$$

which generalizes the divide-and-conquer calculation method we discussed in Chapter 1.

(c) The divide-and-conquer calculation method in (b) above can be extended to compute expected values using the total expectation theorem:

$$\mathbf{E}[X] = \sum_y p_Y(y)\mathbf{E}[X \mid Y = y].$$

The concepts and methods of this chapter extend appropriately to general random variables (see the next chapter), and are fundamental for our subject.