4 Expectation

In Chapter 3, we studied several common discrete distributions. In this chapter we will learn how to obtain their mean, or expectation. We will also cover some useful tools that help us to simplify deriving expectations, such as the linearity of expectation result and deriving expectations by conditioning.

4.1 Expectation of a Discrete Random Variable

The probability mass function (p.m.f.) of a random variable (r.v.) specifies the possible values of the r.v., each with a probability ("weight"). The *expectation* of the random variable, also known as its *mean* or *average*, is a way of summarizing all these different values into a single number. This single number is the sum of all the values, each weighted by its probability of occurring. Expectation is typically used to give us a single value when trading off different options.

Example 4.1 (Choosing between startups)

Suppose you have to choose between startups to join. Startup A will give you a win of ten million dollars with probability 10%, but will cost you one million dollars with probability 90%. Startup B will give you a win of one million dollars with probability 50%, but will cost you half a million with probability 50%.

Question: Which do you choose?

Answer: One way of comparing the two options is to think of A and B as random variables and compare their expectations:

$$\text{Expected value of A } = 10^7 \cdot (0.1) + (-10^6) \cdot (0.9) = 10^5.$$
$$\text{Expected value of B } = 10^6 \cdot (0.5) + (-0.5 \cdot 10^6) \cdot (0.5) = 2.5 \cdot 10^5.$$

By this metric, one might choose startup B. On the other hand, one could also say that expectation is not the right view, since no startup is worth joining if there isn't a potential upside of at least 10 million dollars.

Definition 4.2 *The* **expectation of a discrete random variable** X*, written* $\mathbf{E}[X]$*, is the sum of the possible values of* X*, each weighted by its probability:*

$$\mathbf{E}[X] = \sum_x x\mathbf{P}\{X = x\}.$$

We can also think of $\mathbf{E}[X]$ *as representing the mean of the distribution from which* X *is drawn.*

The following example illustrates why expectation is thought of as an average.

Example 4.3 (Average cost of lunch)

Table 4.1 shows the daily cost of my lunch. What is the average cost of my lunch?

Mon	Tues	Wed	Thurs	Fri	Sat	Sun
\$7	\$7	\$12	\$12	\$12	\$0	\$9

Table 4.1 *Cost of lunch example.*

We can think of *Cost* as a r.v. that takes on each of the values in Table 4.1 with probability $\frac{1}{7}$. Then,

$$\boxed{\text{Average Cost} = \frac{7+7+12+12+12+0+9}{7}}$$

$$|||$$

$$\boxed{\mathbf{E}[\text{Cost}] = 7\cdot\left(\frac{2}{7}\right) + 12\cdot\left(\frac{3}{7}\right) + 9\cdot\left(\frac{1}{7}\right) + 0\cdot\left(\frac{1}{7}\right).}$$

In the expectation view, each possible value (7, 12, 9, and 0) is weighted by its probability.

Question: If $X \sim \text{Bernoulli}(p)$, what is $\mathbf{E}[X]$?

Answer: $\mathbf{E}[X] = 0\cdot(1-p) + 1\cdot(p) = p$.

Example 4.4 (Expected time until disk fails)

Question: Suppose a disk has probability $\frac{1}{3}$ of failing each year. On average, how many years will it be until the disk fails?

Answer: This is simply $\mathbf{E}[X]$, where $X \sim \text{Geometric}(p)$, with $p = \frac{1}{3}$. Assuming $X \sim \text{Geometric}(p)$, we have:

$$\begin{aligned}
\mathbf{E}[X] &= \sum_{n=1}^{\infty} n(1-p)^{n-1}p \\
&= p \cdot \sum_{n=1}^{\infty} n \cdot q^{n-1} \qquad \text{where } q = (1-p) \\
&= p \cdot \left(1 + 2q + 3q^2 + 4q^3 + \ldots\right) \\
&= p \cdot \frac{1}{(1-q)^2} \qquad \text{using (1.4)} \\
&= p \cdot \frac{1}{p^2} \\
&= \frac{1}{p}.
\end{aligned}$$

So when $p = \frac{1}{3}$, the expected number of years until the disk fails is 3. (This type of analysis will be repeated throughout the book, so commit it to memory.)

Question: If $X \sim \text{Poisson}(\lambda)$, what is $\mathbf{E}[X]$?

Answer:

$$\begin{aligned}
\mathbf{E}[X] &= \sum_{i=0}^{\infty} i\frac{e^{-\lambda}\lambda^i}{i!} \\
&= \sum_{i=1}^{\infty} i\frac{e^{-\lambda}\lambda^i}{i!} \\
&= \lambda e^{-\lambda} \sum_{i=1}^{\infty} \frac{\lambda^{i-1}}{(i-1)!} \\
&= \lambda e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \\
&= \lambda e^{-\lambda} e^{\lambda} \qquad \text{using (1.11)} \\
&= \lambda.
\end{aligned}$$

It is interesting to note that the λ parameter for the Poisson distribution is also its mean. The same holds for the p parameter of the Bernoulli distribution. By contrast, the p parameter of the Geometric distribution is the reciprocal of its mean.

One can also consider the expectation of a function of a random variable.

Definition 4.5 *The* **expectation of a function** $g(\cdot)$ *of a discrete random variable* X *is defined as follows:*

$$\mathbf{E}\left[g(X)\right] = \sum_{x} g(x) \cdot p_X(x).$$

Example 4.6 (Volume of sphere)

Consider a sphere, where the radius is a random variable, R, where

$$R = \begin{cases} 1 & \text{w/ prob. } \frac{1}{3} \\ 2 & \text{w/ prob. } \frac{1}{3} \\ 3 & \text{w/ prob. } \frac{1}{3} \end{cases}.$$

Question: What is the expected volume of the sphere?

Answer:

$$\begin{aligned} \mathbf{E}\left[\text{Volume}\right] &= \mathbf{E}\left[\frac{4}{3}\pi R^3\right] \\ &= \frac{4}{3}\pi \cdot 1^3 \cdot \frac{1}{3} + \frac{4}{3}\pi \cdot 2^3 \cdot \frac{1}{3} + \frac{4}{3}\pi \cdot 3^3 \cdot \frac{1}{3} \\ &= 16\pi. \end{aligned}$$

Observe that

$$\mathbf{E}\left[R^3\right] \neq \left(\mathbf{E}\left[R\right]\right)^3.$$

Question: Suppose X is defined as follows:

$$X = \begin{cases} 0 & \text{w/ prob. } 0.2 \\ 1 & \text{w/ prob. } 0.5 \\ 2 & \text{w/ prob. } 0.3 \end{cases}.$$

What is $\mathbf{E}\left[X\right]$ and what is $\mathbf{E}\left[2X^2+3\right]$?

Answer:

$$\begin{aligned} \mathbf{E}\left[X\right] &= 0 \cdot (0.2) + 1 \cdot (0.5) + 2 \cdot (0.3). \\ \mathbf{E}\left[2X^2+3\right] &= \Big(2 \cdot 0^2 + 3\Big)(0.2) + \Big(2 \cdot 1^2 + 3\Big)(0.5) + \Big(2 \cdot 2^2 + 3\Big)(0.3). \end{aligned}$$

You may have noticed that $\mathbf{E}\left[2X^2+3\right] = 2\mathbf{E}\left[X^2\right] + 3$. This is no coincidence and is due to Linearity of Expectation, to be discussed in Section 4.2.

We can also consider the expectation of a function of multiple random variables.

Definition 4.7 *Let X and Y be random variables. The* **expectation of the product** *XY is defined by summing over all possible outcomes (x, y) as follows:*

$$\mathbf{E}[XY] = \sum_x \sum_y xy \cdot p_{X,Y}(x, y),$$

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

Theorem 4.8 (Expectation of a product) *If $X \perp Y$, then*

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

Proof:

$$\begin{aligned}
\mathbf{E}[XY] &= \sum_x \sum_y xy \cdot \mathbf{P}\{X = x, Y = y\} \\
&= \sum_x \sum_y xy \cdot \mathbf{P}\{X = x\}\,\mathbf{P}\{Y = y\} \quad \text{(by definition of } \perp\text{)} \\
&= \sum_x x\mathbf{P}\{X = x\} \cdot \sum_y y\mathbf{P}\{Y = y\} \\
&= \mathbf{E}[X]\,\mathbf{E}[Y].
\end{aligned}$$

■

The same proof shows that if $X \perp Y$, then

$$\mathbf{E}[g(X)f(Y)] = \mathbf{E}[g(X)] \cdot \mathbf{E}[f(Y)], \tag{4.1}$$

for arbitrary functions g and f. A consequence of (4.1) is that if $X \perp Y$, then:

$$\mathbf{E}\left[\frac{X}{Y}\right] = \mathbf{E}[X] \cdot \mathbf{E}\left[\frac{1}{Y}\right].$$

Question: If $\mathbf{E}[XY] = \mathbf{E}[X]\,\mathbf{E}[Y]$, does that imply that $X \perp Y$?

Answer: No, see Exercise 4.7.

We end this section with Theorem 4.9, which offers an alternative way of computing expectations that can be very useful in practice. *Remember this!*

Theorem 4.9 (Alternative definition of expectation) *Let r.v. X be non-negative, discrete, and integer-valued. Then*

$$\mathbf{E}[X] = \sum_{x=0}^{\infty} \mathbf{P}\{X > x\}. \tag{4.2}$$

Proof: See Exercise 4.16. ■

4.2 Linearity of Expectation

The following is one of the most powerful theorems of probability:

Theorem 4.10 (Linearity of Expectation) *For random variables X and Y,*

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

Question: Does Theorem 4.10 require $X \perp Y$?

Answer: Surprisingly not!

Proof: Theorem 4.10 holds for both discrete and continuous random variables. We show below a proof for the case of discrete random variables and will re-prove this in Chapter 8 for the case of continuous random variables.

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

■

Observe that the same proof can also be used to show that

$$\mathbf{E}[f(X) + g(Y)] = \mathbf{E}[f(X)] + \mathbf{E}[g(Y)].$$

Linearity of Expectation can simplify many proofs. We show some examples.

Example 4.11 (Mean of Binomial)

Let $X \sim \text{Binomial}(n, p)$. What is $\mathbf{E}[X]$?

Recall $\mathbf{E}[X] = \sum_{i=0}^{n} i \binom{n}{i} p^i (1-p)^{n-i}$. This expression may appear daunting.

Question: Can we instead think of $\text{Binomial}(n, p)$ as a sum of random variables?

Answer:

$$X = \text{number of heads (successes) in } n \text{ trials} = X_1 + X_2 + \cdots + X_n,$$

where

$$X_i = \begin{cases} 1 & \text{if trial } i \text{ is successful} \\ 0 & \text{otherwise} \end{cases}$$
$$\mathbf{E}[X_i] = p.$$

Then,

$$\mathbf{E}[X] = \mathbf{E}[X_1] + \mathbf{E}[X_2] + \cdots + \mathbf{E}[X_n] = n\mathbf{E}[X_i] = np.$$

Question: What is the intuition behind this result?

Answer: There are n coin flips, each with probability p of coming up heads, which should result in an average of np heads.

The X_i's above are called **indicator random variables** because they take on values 0 or 1. In the previous example, the X_i's were **independent and identically distributed** (**i.i.d.**). However, even if the trials were *not* independent, we would still have

$$\mathbf{E}[X] = \mathbf{E}[X_1] + \cdots + \mathbf{E}[X_n].$$

The following example makes this clear.

Example 4.12 (Drinking from your own cup)

At a party, n people put their drink on a table. Later that night, no one can remember which cup is theirs, so they simply each grab any cup at random (Figure 4.1). Let X denote the number of people who get back their own drink. Think of this as a random permutation of cups across people.

Question: What is $\mathbf{E}[X]$? How do you imagine that $\mathbf{E}[X]$ might depend on n?

Hint: Start by trying to express X as a sum of indicator random variables?

Answer: $X = I_1 + I_2 + \cdots + I_n$, where

$$I_i = \begin{cases} 1 & \text{if the } i\text{th person gets their own drink} \\ 0 & \text{otherwise} \end{cases}.$$

Although the I_i's have the same distribution (by symmetry), they are *not* independent of each other! Nevertheless, we can still use Linearity of Expectation to

Figure 4.1 *Each person picks up a random cup.*

say

$$\begin{aligned}\mathbf{E}[X] &= \mathbf{E}[I_1] + \mathbf{E}[I_2] + \cdots + \mathbf{E}[I_n] \\ &= n\mathbf{E}[I_i] \\ &= n\left(\frac{1}{n} \cdot 1 + \frac{n-1}{n} \cdot 0\right) \\ &= 1.\end{aligned}$$

Interestingly, the expected number of people who get back their own drink is independent of n!

Example 4.13 (Coupon collector)

Imagine there are n distinct coupons that we are trying to collect (Figure 4.2). Every time that we draw a coupon, we get one of the n coupons at *random*, with each coupon being equally likely. (You can think of this as draws with replacement, or you can imagine that there are an infinite number of each of the n coupon types.) Thus it is quite likely that the same coupon will be drawn more than one time. The coupon collector question asks:

How many draws does it take in expectation until I get all n distinct coupons?

Let D denote the number of draws needed to collect all coupons.

Question: What is $\mathbf{E}[D]$?

Answer: It is not at all obvious how to get $\mathbf{E}[D]$. The trick is to try to express

Figure 4.2 *The goal of the coupon collector problem is to collect all n coupons.*

D as a sum of random variables:

$$D = D_1 + D_2 + \cdots + D_n. \tag{4.3}$$

Question: What should D_i represent?

Answer: One might think that D_i should be the number of draws needed to get coupon number i. But this doesn't work, because while I'm trying to get coupon i, I might be drawing other coupons.

Question: Is there a better definition for D_i that doesn't result in over-counting?

Answer: Let D_i denote the number of draws needed to get the ith *distinct* coupon, after getting $i-1$ distinct coupons. That is, D_1 is the number of draws needed to get any coupon (namely $D_1 = 1$). D_2 is the number of *additional* draws needed to get a coupon which is distinct from the first coupon. D_3 is the number of *additional* draws needed to get a coupon which is distinct from the first two distinct coupons.

Question: How is D_i distributed?

Answer:

$$\begin{aligned}
D_1 &\sim \text{Geometric}\,(1) = 1 \\
D_2 &\sim \text{Geometric}\left(\frac{n-1}{n}\right) \\
D_3 &\sim \text{Geometric}\left(\frac{n-2}{n}\right) \\
&\vdots \\
D_n &\sim \text{Geometric}\left(\frac{1}{n}\right).
\end{aligned}$$

We are now finally ready to apply Linearity of Expectation to (4.3).

$$\begin{aligned}\mathbf{E}[D] &= \mathbf{E}[D_1 + D_2 + \cdots + D_n] \\ &= \mathbf{E}[D_1] + \mathbf{E}[D_2] + \cdots + \mathbf{E}[D_n] \\ &= 1 + \frac{n}{n-1} + \frac{n}{n-2} + \cdots + \frac{n}{1}.\end{aligned}$$

But we can express this in terms of the harmonic series (see Section 1.4) as follows:

$$\mathbf{E}[D] = n \cdot \left(\frac{1}{n} + \frac{1}{n-1} + \frac{1}{n-2} + \cdots + 1\right) = n \cdot H_n, \tag{4.4}$$

where

$$H_n = 1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n}.$$

Question: What is $\mathbf{E}[D]$ approximately equal to for large n?

Answer: From (1.17), it follows that $\mathbf{E}[D] \approx n \ln n$.

4.3 Conditional Expectation

One is often interested in the expected value of a random variable conditioned on some event. For example, if X is a random variable denoting the price of a hotel room and A is the event that the month is March, one might be interested in $\mathbf{E}[X \mid A]$, which is the expected price of the room given that the month is March.

Recall that $p_X(\cdot)$ is the p.m.f. for r.v. X, where

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

To understand conditional expectation, rather than working with the p.m.f., we need to work with the *conditional p.m.f.*

Definition 4.14 *Let X be a discrete r.v. with p.m.f. $p_X(\cdot)$ defined over a countable sample space. Let A be an event s.t. $\mathbf{P}\{A\} > 0$. Then $p_{X|A}(\cdot)$ is the* **conditional p.m.f.** *of X given event A. We define*

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

A conditional probability thus involves narrowing down the probability space. To see this, let's consider some examples.

Example 4.15 (Conditioning on an event)

Let X denote the size of a job. Suppose that

$$X = \begin{cases} 1 & \text{w/ prob. } 0.1 \\ 2 & \text{w/ prob. } 0.2 \\ 3 & \text{w/ prob. } 0.3 \\ 4 & \text{w/ prob. } 0.2 \\ 5 & \text{w/ prob. } 0.2 \end{cases}.$$

Let A be the event that the job is "small," meaning that its size is ≤ 3. Our goal is to understand the conditional p.m.f. of X given event A, which is colored in blue.

Question: What is $p_X(1)$?

Answer: $p_X(1) = 0.1$.

Question: What is $p_{X|A}(1)$?

Answer: Intuitively, if we condition on the job being small (blue), we can see that, of the blue jobs, one-sixth of them have size 1. Algebraically:

$$\begin{aligned} p_{X|A}(1) = \mathbf{P}\{X = 1 \mid A\} &= \frac{\mathbf{P}\{X = 1 \ \& \ A\}}{\mathbf{P}\{A\}} \\ &= \frac{\mathbf{P}\{X = 1\}}{\mathbf{P}\{A\}} \\ &= \frac{\frac{1}{10}}{\frac{6}{10}} \\ &= \frac{1}{6}. \end{aligned}$$

We have normalized $\mathbf{P}\{X = 1\}$ by the probability of being in A.

Question: What is $p_{X|A}(x)$, if $x \notin A$?

Answer: 0.

Lemma 4.16 *A conditional p.m.f. is a p.m.f., that is,*

$$\sum_x p_{X|A}(x) = \sum_{x \in A} p_{X|A}(x) = 1.$$

Proof: See Exercise 4.12. ∎

We can also consider the case where the event, A, is an instance of a r.v. For example, A might be the event $Y = y$.

Example 4.17 (Conditioning on the value of a random variable)

Two discrete random variables X and Y taking the values $\{0, 1, 2\}$ have a joint p.m.f. given by Table 4.2.

$Y = 2$	0	$\frac{1}{6}$	$\frac{1}{8}$
$Y = 1$	$\frac{1}{8}$	$\frac{1}{6}$	$\frac{1}{8}$
$Y = 0$	$\frac{1}{6}$	$\frac{1}{8}$	0
	$X = 0$	$X = 1$	$X = 2$

Table 4.2 *Joint p.m.f.,* $p_{X,Y}(x, y)$.

Question: What is $p_{X|Y=2}(1)$?

Answer:

$$p_{X|Y=2}(1) = \mathbf{P}\{X = 1 \mid Y = 2\} = \frac{\mathbf{P}\{X = 1 \ \& \ Y = 2\}}{\mathbf{P}\{Y = 2\}} = \frac{\frac{1}{6}}{\frac{1}{6} + \frac{1}{8}} = \frac{4}{7}.$$

Question: What is $p_{X|Y=2}(2)$?

Answer: By the fact that $p_{X|Y=2}(x)$ is a p.m.f., and observing that $p_{X|Y=2}(0) = 0$, it must be the case that

$$p_{X|Y=2}(2) = 1 - p_{X|Y=2}(1) = \frac{3}{7}.$$

Definition 4.18 *For a discrete r.v. X, the* **conditional expectation** *of X given event A is as follows:*

$$\mathbf{E}[X \mid A] = \sum_x x p_{X|A}(x) = \sum_x x \cdot \frac{\mathbf{P}\{(X = x) \cap A\}}{\mathbf{P}\{A\}}.$$

Simply put, the conditional expectation is the same as the expectation, but rather than using the p.m.f., we use the conditional p.m.f., which likely has a different range.

Example 4.19 (Conditional expectation)

Again let X denote the size of a job. Suppose that

$$X = \begin{cases} 1 & \text{w/ prob. } 0.1 \\ 2 & \text{w/ prob. } 0.2 \\ 3 & \text{w/ prob. } 0.3 \\ 4 & \text{w/ prob. } 0.2 \\ 5 & \text{w/ prob. } 0.2 \end{cases}.$$

Let A be the event that the job is "small," meaning that its size is ≤ 3.

Question: What is $\mathbf{E}[X]$?

Answer:

$$\mathbf{E}[X] = 1 \cdot \frac{1}{10} + 2 \cdot \frac{2}{10} + 3 \cdot \frac{3}{10} + 4 \cdot \frac{2}{10} + 5 \cdot \frac{2}{10} = \frac{32}{10}.$$

Question: What is $\mathbf{E}[X|A]$?

Answer: Note that this should be a smaller value than $\mathbf{E}[X]$.

$$\begin{aligned} \mathbf{E}[X|A] &= 1 \cdot p_{X|A}(1) + 2 \cdot p_{X|A}(2) + 3 \cdot p_{X|A}(3) \\ &= 1 \cdot \frac{1}{6} + 2 \cdot \frac{2}{6} + 3 \cdot \frac{3}{6} \\ &= \frac{14}{6}. \end{aligned}$$

Example 4.20 (More conditional expectation practice)

Two discrete random variables X and Y taking the values $\{0, 1, 2\}$ have a joint p.m.f. given by Table 4.3.

$Y = 2$	0	$\frac{1}{6}$	$\frac{1}{8}$
$Y = 1$	$\frac{1}{8}$	$\frac{1}{6}$	$\frac{1}{8}$
$Y = 0$	$\frac{1}{6}$	$\frac{1}{8}$	0
	$X = 0$	$X = 1$	$X = 2$

Table 4.3 *Joint p.m.f.,* $p_{X,Y}(x, y)$.

Question: Compute the conditional expectation $\mathbf{E}[X \mid Y = 2]$.

Answer:

$$\begin{aligned}\mathbf{E}[X \mid Y=2] &= 0 \cdot p_{X|Y=2}(0) + 1 \cdot p_{X|Y=2}(1) + 2 \cdot p_{X|Y=2}(2) \\ &= 1 \cdot \frac{4}{7} + 2 \cdot \frac{3}{7} = \frac{10}{7}.\end{aligned}$$

Example 4.21 (Indicators: an alternative to conditioning)

Let S be a discrete r.v., without loss of generality, say:

$$S = \begin{cases} 1 & \text{w/prob } p_S(1) \\ 2 & \text{w/prob } p_S(2) \\ 3 & \text{w/prob } p_S(3) \\ 4 & \text{w/prob } p_S(4) \\ & \vdots \end{cases}.$$

Let $I_{S \le x}$ be an indicator r.v. which is 1 when $S \le x$ and 0 otherwise. Likewise, let $I_{S>x}$ be an indicator r.v. which is 1 when $S > x$ and 0 otherwise.

Question: Argue that

$$S \overset{d}{=} S \cdot I_{S \le x} + S \cdot I_{S>x}. \tag{4.5}$$

The $\overset{d}{=}$ is indicating that the left-hand side and right-hand side of (4.5) are **equal in distribution**, that is, they take on the same values with the same probabilities.

Answer: $S \cdot I_{S \le x}$ is a r.v. that returns the same values of S if those values are $\le x$, and otherwise returns 0. Think of this as the r.v. S with a bunch of 0's where the terms for $S > x$ would be. For example, if $x = 2$, then:

$$S \cdot I_{S \le 2} = \begin{cases} 1 & \text{w/prob } p_S(1) \\ 2 & \text{w/prob } p_S(2) \\ 0 & \text{w/prob } p_S(3) \\ 0 & \text{w/prob } p_S(4) \\ & \vdots \end{cases} \qquad S \cdot I_{S>2} = \begin{cases} 0 & \text{w/prob } p_S(1) \\ 0 & \text{w/prob } p_S(2) \\ 3 & \text{w/prob } p_S(3) \\ 4 & \text{w/prob } p_S(4) \\ & \vdots \end{cases}.$$

Adding together $S \cdot I_{S \le x}$ and $S \cdot I_{S>x}$, we get exactly the distribution of S.

Question: How does $S \cdot I_{S \le 2}$ compare to the r.v. $[S \mid S \le 2]$?

Answer:

$$S \cdot I_{S \le 2} = \begin{cases} 1 & \text{w/prob } p_S(1) \\ 2 & \text{w/prob } p_S(2) \\ 0 & \text{w/prob } 1 - \mathbf{P}\{S \le 2\} \end{cases}.$$

By contrast,

$$[S \mid S \leq 2] = \begin{cases} 1 & \text{w/prob } p_S(1)/\mathbf{P}\{S \leq 2\} \\ 2 & \text{w/prob } p_S(2)/\mathbf{P}\{S \leq 2\} \end{cases}.$$

Question: How is $\mathbf{E}[S \cdot I_{S\leq 2}]$ related to $\mathbf{E}[S \mid S \leq 2]$?

Answer:

$$\mathbf{E}[S \cdot I_{S\leq 2}] = 1 \cdot p_S(1) + 2 \cdot p_S(2).$$

More generally,

$$\mathbf{E}[S \cdot I_{S\leq x}] = \sum_{i=1}^{x} i p_S(i). \tag{4.6}$$

By contrast,

$$\mathbf{E}[S \mid S \leq x] = \sum_{i=1}^{x} i \frac{p_S(i)}{\mathbf{P}\{S \leq x\}} = \frac{1}{\mathbf{P}\{S \leq x\}} \cdot \sum_{i=1}^{x} i p_S(i). \tag{4.7}$$

Comparing (4.6) and (4.7), we have:

$$\mathbf{E}[S \cdot I_{S\leq x}] = \mathbf{E}[S \mid S \leq x] \cdot \mathbf{P}\{S \leq x\}. \tag{4.8}$$

Question: Express $\mathbf{E}[S]$ in two ways: (1) using indicator random variables and (2) via conditioning on $S \leq x$.

Answer: For (1), we use (4.5) and take expectations of both sides as follows:

$$\mathbf{E}[S] = \mathbf{E}[S \cdot I_{S\leq x}] + \mathbf{E}[S \cdot I_{S>x}]. \tag{4.9}$$

For (2) we use the result from (4.8) to replace each term in (4.9), obtaining:

$$\mathbf{E}[S] = \mathbf{E}[S \mid S \leq x] \cdot \mathbf{P}\{S \leq x\} + \mathbf{E}[S \mid S > x] \cdot \mathbf{P}\{S > x\}. \tag{4.10}$$

4.4 Computing Expectations via Conditioning

Recall the Law of Total Probability, which says that the probability of an event can be computed as a sum of conditional probabilities. In the same way, an expectation can be computed as a sum of conditional expectations – we saw an example of this in (4.10). Conditioning is often the easiest way to compute an expectation.

Theorem 4.22 *Let events $F_1, F_2, F_3, \ldots$ partition the sample space Ω. Then,*

$$\mathbf{E}[X] = \sum_{i=1}^{\infty} \mathbf{E}[X \mid F_i] \cdot \mathbf{P}\{F_i\}.$$

Given a discrete r.v. Y, if we think of $Y = y$ as an event, then we have:

$$\mathbf{E}[X] = \sum_{y} \mathbf{E}[X \mid Y = y] \mathbf{P}\{Y = y\}.$$

Proof: We show the proof for the second expression in the theorem. The proof for the first expression follows the same lines.

$$\begin{aligned}
\mathbf{E}[X] &= \sum_x x\mathbf{P}\{X = x\} \\
&= \sum_x x \sum_y \mathbf{P}\{X = x \mid Y = y\} \mathbf{P}\{Y = y\} \\
&= \sum_x \sum_y x\mathbf{P}\{X = x \mid Y = y\} \mathbf{P}\{Y = y\} \\
&= \sum_y \sum_x x\mathbf{P}\{X = x \mid Y = y\} \mathbf{P}\{Y = y\} \\
&= \sum_y \mathbf{P}\{Y = y\} \sum_x x\mathbf{P}\{X = x \mid Y = y\} \\
&= \sum_y \mathbf{P}\{Y = y\} \mathbf{E}[X \mid Y = y] \quad \text{(by Definition 4.18).}
\end{aligned}$$

■

Example 4.23 (Expectation of Geometric, revisited)

Recall that in Example 4.4 we computed the mean of a Geometric with parameter p. How can we redo this more simply via conditioning? Specifically, we seek $\mathbf{E}[N]$, where N is the number of flips required to get the first head.

Question: What do we condition on?

Answer: We condition on the value of the first flip, Y, as follows:

$$\begin{aligned}
\mathbf{E}[N] &= \mathbf{E}[N \mid Y = 1] \mathbf{P}\{Y = 1\} + \mathbf{E}[N \mid Y = 0] \mathbf{P}\{Y = 0\} \\
&= 1 \cdot p + (1 + \mathbf{E}[N]) \cdot (1 - p) \\
p\mathbf{E}[N] &= p + (1 - p) \\
\mathbf{E}[N] &= \frac{1}{p}.
\end{aligned}$$

The difficult step here is reasoning that $\mathbf{E}[N \mid Y = 0] = 1 + \mathbf{E}[N]$. That is,

knowing that we already got a tail on the first flip adds 1 to the expected time to get a head, because the *remaining* time needed to get a head "restarts" after that tail. This is the same idea as a person who has been trying to win the lottery for the last 100 days. Their remaining time to win the lottery is the same as if they started today. The fact that they already tried for 100 days just adds 100 to their total time spent trying to win the lottery. The property that your past doesn't affect your future is called **memorylessness** and will come up again.

Note how conditioning greatly simplifies the original derivation given in Example 4.4.

The proof of Theorem 4.22 generalizes to Theorem 4.24:

Theorem 4.24

$$\mathbf{E}\left[g(X)\right] = \sum_{y} \mathbf{E}\left[g(X) \mid Y = y\right] \mathbf{P}\left\{Y = y\right\}.$$

Theorem 4.24 will be particularly useful in Chapter 5's discussion of higher moments.

4.5 Simpson's Paradox

We end this chapter with Simpson's paradox [70]. The paradox is counter-intuitive because people *mistakenly* think it is related to conditioning, when it is not.

The best way to understand Simpson's paradox is via an example. A common example from the healthcare area involves the evaluation of two potential treatments for kidney stones: call these Treatment A and Treatment B. Suppose that patients are classified as having "small" kidney stones or "large" ones. It turns out that Treatment A is more effective than B on small stones, and also that Treatment A is more effective than B on large stones. However, paradoxically, if we ignore patient classifications, we find that Treatment B is the more effective treatment. The fact that the "winner" changes when we remove the classification is called Simpson's paradox.

Question: Spend some time asking yourself: How can this be?

Answer: Table 4.4 shows a numerical instance of the paradox. Looking at the top left box, (small, A), we see that Treatment A is 90% effective on small stones – it is effective on 90 out of the 100 small-stone patients who receive

	Treatment A	Treatment B
Small stones	90% effective **(winner!)** (successful on 90 out of 100)	80% effective (successful on 800 out of 1000)
Large stones	60% effective **(winner!)** (successful on 600 out of 1000)	50% effective (successful on 50 out of 100)
Aggregate mix	63% effective (successful on 690 out of 1100)	77% effective **(winner!)** (successful on 850 out of 1100)

Table 4.4 *Simpson's paradox: Treatment A is more effective than Treatment B both on small stones and on large stones. But Treatment B is more effective than Treatment A when we ignore stone size.*

it. By contrast, Treatment B is only 80% effective on small stones, as shown in box (small, B) – it is effective on 800 out of the 1000 small-stone patients who receive it. Thus Treatment A is more effective than Treatment B on small-stone patients. The large stone row of the table shows that Treatment A is 60% effective on large-stone patients, while Treatment B is only 50% effective on large-stone patients. Based on the above data, it seems that Treatment A is best.

In the last line of the table, labeled "aggregate mix," we mix up all the small-stone and large-stone patients, so that they are no longer classified by their stone size. We now look at the 1100 patients that received Treatment A and ask how many of them had success. We find that only 690 of the 1100 patients had success, meaning that Treatment A is 63% effective. By contrast, of the 1100 patients that received Treatment B, we find that 77% of them had success. Based on this, it seems that Treatment B is best.

Question: Which treatment is actually best, A or B?

Answer: Treatment A is best. Treatment A is best when used for patients with small stones, and it is also best when used for patients with large stones. In practice, doctors know that Treatment A is best, and they thus reserve it for patients with large stones, which are the more difficult cases. This is why we see bigger studies (1000 patients) where Treatment A is applied to patients with large stones. Treatment B is more typically reserved for the easier patients, which is why we see bigger studies (1000 patients) where Treatment B is applied to patients with small stones.

Question: But if Treatment A is best, why does it turn out to look bad in the "mix," where we ignore the patient classification?

Answer: Mathematically, the paradox is caused by a combination of two things:

1. The biggest contributors to the "mix" are quadrants [large, A] and [small, B], since these both involve tests with 1000 patients.
2. But [small, B] has a higher effectiveness percentage than [large, A] because, although Treatment A is the better treatment, this fact is dwarfed by the fact that small stones are so much easier to handle than large ones.

Together, these leave us believing that Treatment B is better when we look at the aggregate mix.

4.6 Exercises

4.1 **Socks**
Socks come in two colors: red and blue. There are an infinite number of socks of each color. Each time we pick a sock, we get a random sock. What is the expected number of picks until we have a pair (two of the same color)?

4.2 **Random graphs**
Consider a "random graph" on n vertices, where each pair of vertices is connected by an edge with probability p.
(a) What is the expected number of edges in the random graph?
(b) What is the distribution of the degree of vertex i?

4.3 **Multi-round gamble**
[Proposed by Rashmi Vinayak] At a casino, you're attracted to an "amazing" offer. Every round you bet, you either triple your bet with probability half or lose your bet with probability half.
(a) What is the obvious betting strategy, that is, how much of your money should you bet in each round to maximize your winnings?
(b) For this strategy, what is the probability that you end up with no money? (Are you surprised by the answers to the above two questions?)

4.4 **Probability bounds**
You are told that the average file size in a database is 6K bytes.
(a) Explain why it follows (from the definition of expectation) that fewer than half of the files can have size > 12K.
(b) You are now given the additional information that the minimum file size is 3K. Derive a tighter upper bound on the percentage of files of size > 12K.

4.5 **Shared birthdays**
There are n people in a room. A person is "happy" if he/she shares a birthday with another person in the room. What is the expected number of happy people?

4.6 **Identities**
Let A and B be independent random variables. Assume that $B \neq 0$ and that $\mathbf{E}[B] \neq 0$. Prove or disprove the following statement:

$$\mathbf{E}\left[\frac{A}{B}\right] = \frac{\mathbf{E}[A]}{\mathbf{E}[B]}.$$

4.7 **Expectation of product**
Prove or disprove the following claim: If

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

then X and Y are independent random variables.

4.8 **Coin pattern**
Your coin produces heads with probability p. You flip your coin $n \geq 3$ times. What is the expected number of times that the pattern HTH appears? Example: If the n flips are $\langle H, T, H, T, H, T, H \rangle$ then the pattern HTH appears three times.

4.9 **Random word generation: I love to love**
Every minute, a random word generator spits out one word uniformly at random from the set { I, love, to }. Suppose we let the generator run for n minutes. What is the expected number of times that the phrase "I love to love" appears?

4.10 **Permutations**
Let π be a permutation on $[n] = \{1, \ldots, n\}$, where $n \geq 3$. Here $\pi(i)$ denotes the number in the ith position of permutation π. We say that π has a local maximum at $i \in [n]$ if all these are true:
- $\pi(i) > \pi(i+1)$, if $i = 1$
- $\pi(i-1) < \pi(i)$ and $\pi(i) > \pi(i+1)$, if $1 < i < n$
- $\pi(i-1) < \pi(i)$, if $i = n$

What is the expected number of local maxima of a random permutation π on $[n]$? [Hint: Use Linearity of Expectation and indicator random variables.]

4.11 **Triangles in random graphs**
Consider a "random graph," G, on n vertices, where each pair of vertices is connected by an edge with probability $p = \frac{d}{n}$. Let Y denote the number of triangles in G. Derive $\mathbf{E}[Y]$. What does $\mathbf{E}[Y]$ look like for high n?

4.12 **A conditional p.m.f. is a p.m.f**
Prove that the conditional p.m.f. $p_{X|A}(\cdot)$ is a p.m.f. by showing that

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

4.13 **Tail of Geometric and memorylessness**
Let $X \sim \text{Geometric}(p)$.
(a) Derive $\mathbf{E}[X \mid X > 5]$ by summing over the conditional p.m.f. Be careful to get the indices correct.
(b) Your final answer should be extremely simple in light of the memorylessness property of the Geometric distribution. Explain your final answer.

4.14 **Practice with conditional expectation**
For the joint p.m.f. in Table 4.2, compute $\mathbf{E}[X \mid Y \neq 1]$.

4.15 **More conditional expectation practice**
We're given a joint p.m.f. for two random variables X and Y.

	$Y=1$	$Y=2$	$Y=3$
$X=0$	1/4	3/16	1/16
$X=1$	1/8	0	3/8

What is $\mathbf{E}\left[\frac{X}{Y} \,\middle|\, X^2+Y^2 \leq 4\right]$?

4.16 **Alternative definition of expectation: summing the tail**
Let X be a non-negative, discrete, integer-valued r.v. Prove that

$$\mathbf{E}[X] = \sum_{x=0}^{\infty} \mathbf{P}\{X > x\}.$$

4.17 **Simpson's paradox for PhD admissions**
A total of 110 Berkeley undergrads and 110 CMU undergrads apply for PhD programs in CS at Berkeley and CMU. Assume that all the students started their grad applications at the very last minute, and as a result were only able to apply to their first choice between Berkeley PhD or CMU PhD (not both). Below are the acceptance rates for each group at each university:

	Berkeley Undergrad	CMU Undergrad
Berkeley PhD	32% (32 out of 100 applicants)	40% (4 out of 10 applicants)
CMU PhD	10% (1 out of 10 applicants)	18% (18 out of 100 applicants)
Either PhD	30% (33 out of 110 applicants)	20% (22 out of 110 applicants)

(a) Which group of students, CMU students or Berkeley students, had a higher acceptance rate into the Berkeley PhD program?
(b) Which group of students, CMU students or Berkeley students, had a higher acceptance rate into the CMU PhD program?
(c) Which group of students, CMU students or Berkeley students, were more likely to be admitted into a PhD program?
(d) What was Berkeley's overall acceptance rate and what was CMU's overall acceptance rate (assume that no students outside Berkeley or CMU applied)?
(e) What proportion of the students admitted to either the CMU or Berkeley PhD programs were admitted to the Berkeley PhD program?
(f) How is it possible that CMU students had a higher acceptance rate at each of the PhD programs than Berkeley students, and yet a lower chance of getting into a PhD program overall?

4.18 ***k* heads in a row**

Stacy's fault-tolerant system only crashes if there are k *consecutive* failures. Assume that every minute a failure occurs independently with probability p. What is the expected number of minutes until Stacy's system crashes? This is equivalent to $\mathbf{E}[T_k]$, where T_k denotes the number of flips needed to get k heads in a row when flipping a coin with probability p of heads. [Hint: Write a recurrence relation for the r.v. T_k in terms of T_{k-1}.]

4.19 **Virus propagation**

We start with a network of three computers. Unbeknownst to us, two of the computers are infected with a hidden virus and the other is not. A sequence of new uninfected computers now join the network, one at a time. Each new computer joins the existing network by attaching itself to a random computer in the network (all computers in the network are equally likely attachment points). If the new computer attaches itself to an infected computer, then it immediately becomes infected with the virus; otherwise the new computer does not get the virus. At the point where the network consists of k total computers, what is the expected fraction of these that is infected? Assume $k > 3$.

4.20 **Coupon collector, time to repeat**

In the coupon collection problem, there are n distinct coupons that we are trying to collect. Every time we draw a coupon, we get one of the n at random, with each coupon being equally likely (the coupon we get is replaced after each drawing). Thus it is likely that we quickly see a repeat. Define N to be the r.v., where

$$N = \text{Number of coupons collected until we first get a repeat.}$$

What is $\mathbf{E}[N]$? [Note: You can leave your answer in the form of a sum.]

4.21 **Minimum of n dice**

Your goal is to derive a simple expression for the expected minimum value of n independent rolls of a die. Below are some steps to help:

(a) You roll a die twice. What's the expected value of the minimum of the rolls? Compute this by simple counting and/or conditioning.

(b) Now redo the problem in (a), but use the result of Exercise 4.16 to compute your answer.

(c) Now repeat (b), but for the case of n rolls. What does your expression for the expected minimum value become as $n \to \infty$?

4.22 **The counter-intuitive nature of conditional independence**

A fair coin is flipped N times, and H heads are obtained (both N and H are random variables in the experiment). Suppose we are told that $H = 5$. What is $\mathbf{E}[N \mid H = 5]$?

(a) If you had to guess, what would you guess is $\mathbf{E}[N \mid H = 5]$?

(b) Write out $\mathbf{E}[N \mid H = 5]$ based on Definition 4.18. Go as far as you can in evaluating the expression. What is missing?

(c) Suppose we're now given a "prior" distribution on N, namely that

$$N = \begin{cases} 10 & \text{w/prob } 0.5 \\ 20 & \text{w/prob } 0.5 \end{cases}.$$

Use this information to finish evaluating $\mathbf{E}[N \mid H = 5]$. You will need a computer for the final evaluation. Is this the answer you originally expected?

4.23 **Revisiting die rolls**

Recall Exercise 2.27, where you have two dice. Die A is a fair die (each of the six numbers is equally likely) and die B is a biased die (the number six comes up with probability $\frac{2}{3}$ and the remaining $\frac{1}{3}$ probability is split evenly across all the other numbers). Kaige picks a die at random and rolls that die three times. Given that the first two rolls are both sixes, what is the expected value of the third roll?

4.24 **Expectation of a product**

[Proposed by Priyatham Bollimpalli] In his spare time, Priyatham likes to sit around multiplying random numbers. After studying probability, he has become interested in the expected value of the product.

(a) Priyatham multiplies two single-digit, non-equal, positive numbers. If each number is equally likely to be picked from $\{1, 2, 3, \ldots, 9\}$ (without replacement), what is the expected value of the product? [Note: This formula may come in handy: $\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}$]

(b) What is the expectation of a product of two two-digit numbers, where all four digits are non-zero, unique, and picked uniformly at random?

For example 45 and 76 are two valid numbers, whereas 45 and 59 are not since 5 is repeated. [Hint: Solution is short.]

(c) Suppose we didn't need to assume the digits were unique. Explain in one line what the answer to part (a) would be now.

4.25 **Socks for multi-footed animals**

Socks come in two colors: red and blue. There are an infinite number of socks of each color. Every time we pick a sock, we get a random sock.

(a) A human has two feet. What is the expected number of picks until we have a pair (sock for each foot) of a single color?

(b) The three-toed sloth needs a sock for each toe. What is the expected number of picks until we have three socks of the same color?

(c) Prof. Veloso's new soccer-playing robot has n feet (more feet helps it win). What is the expected number of picks until the robot has a sock for each foot, where all socks must be of the same color? [Note: Don't worry about trying to come up with a closed-form expression.]

4.26 **The three rivers of Pittsburgh**

[Proposed by Lea Herzberg] In Pittsburgh, three rivers meet as shown in Figure 4.3. Assume that, for every boat traveling up the Ohio River, with probability $\frac{2}{3}$ the boat continues up the Allegheny and, independently, with probability $\frac{1}{3}$, the boat continues up the Monongahela.

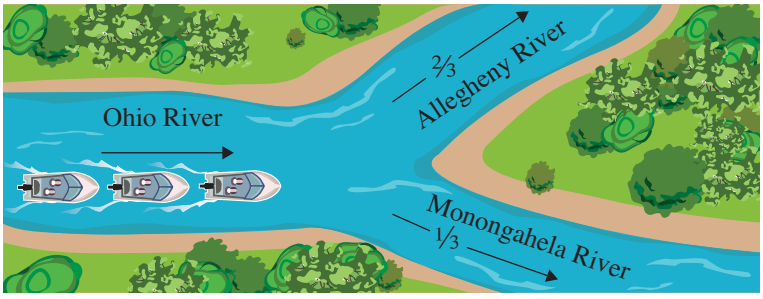

Figure 4.3 *The three rivers of Pittsburgh, where the arrows represent the direction of the boats.*

Let X denote the number of boats approaching the fork from the Ohio in the last hour. Let A (respectively, M) denote the number of boats entering the Allegheny (respectively, Monongahela) in the last hour.

Suppose $X \sim \text{Poisson}(\lambda = 100)$. Your goal is to derive $\mathbf{E}[X \mid M = 100]$.

(a) Do you have any intuition about what $\mathbf{E}[X \mid M = 100]$ should be?

(b) Using Definition 4.18 for conditional expectation, write an expression for $\mathbf{E}[X \mid M = 100]$ using all the information given in the problem. Express your answer in terms of an expression involving λ's and x's and some sums. Do not worry about simplifying your expression.

(c) Your expression in (b) is very unwieldy and hard to evaluate. Instead, we will follow a different approach to get to the answer. In following this approach, assume $p = \frac{1}{3}$ and $\lambda = 100$, but express your answers generally in terms of p and λ until the last part.
   (i) Let $Z = [M \mid X = x]$. How is Z distributed?
   (ii) Using step (i), what is the joint probability $p_{X,M}(x, m)$?
   (iii) Use step (ii) to prove that $M \sim \text{Poisson}(\lambda p)$.
   (iv) Combine steps (ii) and (iii) to derive $p_{X|M=m}(x)$, and then use that to get the distribution of the r.v. $[X - m \mid M = m]$.
   (v) Use the result in (iv) to get $\mathbf{E}[X \mid M = m]$.
   (vi) Returning to the original problem, given that in the last hour 100 boats entered the Monongahela, what is the expected number of boats leaving the Ohio in the last hour? Note that you will likely find your intuition from part (a) was incorrect.