

## 18 Expectation

### 18.1 Definitions and Examples

The *expectation* or *expected value* of a random variable is a single number that tells you a lot about the behavior of the variable. Roughly, the expectation is the average value of the random variable where each value is weighted according to its probability. Formally, the expected value (also known as the *average* or *mean*) of a random variable is defined as follows.

**Definition 18.1.1.** If  $R$  is a random variable defined on a sample space  $\mathcal{S}$ , then the expectation of  $R$  is

$$\text{Ex}[R] ::= \sum_{w \in \mathcal{S}} R(w) \Pr[w]. \quad (18.1)$$

For example, suppose  $\mathcal{S}$  is the set of students in a class, and we select a student uniformly at random. Let  $R$  be the selected student’s exam score. Then  $\text{Ex}[R]$  is just the class average—the first thing everyone wants to know after getting their test back! For similar reasons, the first thing you usually want to know about a random variable is its expected value.

Let’s work through some examples.

#### 18.1.1 The Expected Value of a Uniform Random Variable

Let  $R$  be the value that comes up with you roll a fair 6-sided die. The the expected value of  $R$  is

$$\text{Ex}[R] = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = \frac{7}{2}.$$

This calculation shows that the name “expected value” is a little misleading; the random variable might *never* actually take on that value. You don’t ever expect to roll a  $3\frac{1}{2}$  on an ordinary die!

Also note that the mean of a random variable is not the same as the *median*. The median is the midpoint of a distribution.

**Definition 18.1.2.** The *median*<sup>1</sup> of a random variable  $R$  is the value  $x \in \text{range}(R)$

<sup>1</sup>Some texts define the median to be the value of  $x \in \text{range}(R)$  for which  $\Pr[R \leq x] < 1/2$  and  $\Pr[R > x] \leq 1/2$ . The difference in definitions is not important.

Chapter 18 Expectation

such that

$$\Pr[R \leq x] \leq \frac{1}{2} \quad \text{and}$$

$$\Pr[R > x] < \frac{1}{2}.$$

In this text, we will not devote much attention to the median. Rather, we will focus on the expected value, which is much more interesting and useful.

Rolling a 6-sided die provides an example of a uniform random variable. In general, if  $R_n$  is a random variable with a uniform distribution on  $\{1, 2, \dots, n\}$ , then

$$\text{Ex}[R_n] = \sum_{i=1}^n i \cdot \frac{1}{n} = \frac{n(n+1)}{2n} = \frac{n+1}{2}.$$

### 18.1.2 The Expected Value of an Indicator Random Variable

The expected value of an indicator random variable for an event is just the probability of that event.

**Lemma 18.1.3.** *If  $I_A$  is the indicator random variable for event  $A$ , then*

$$\text{Ex}[I_A] = \Pr[A].$$

*Proof.*

$$\begin{aligned} \text{Ex}[I_A] &= 1 \cdot \Pr[I_A = 1] + 0 \cdot \Pr[I_A = 0] \\ &= \Pr[I_A = 1] \\ &= \Pr[A]. \quad (\text{def of } I_A) \quad \blacksquare \end{aligned}$$

For example, if  $A$  is the event that a coin with bias  $p$  comes up heads, then  $\text{Ex}[I_A] = \Pr[I_A = 1] = p$ .

### 18.1.3 Alternate Definitions

There are several equivalent ways to define expectation.

**Theorem 18.1.4.** *If  $R$  is a random variable defined on a sample space  $S$  then*

$$\text{Ex}[R] = \sum_{x \in \text{range}(R)} x \cdot \Pr[R = x]. \quad (18.2)$$

The proof of Theorem 18.1.4, like many of the elementary proofs about expectation in this chapter, follows by judicious regrouping of terms in the Equation 18.1:

18.1. Definitions and Examples

*Proof.*

$$\begin{aligned}
 \text{Ex}[R] &= \sum_{\omega \in \mathcal{S}} R(\omega) \Pr[\omega] && \text{(Def 18.1.1 of expectation)} \\
 &= \sum_{x \in \text{range}(R)} \sum_{\omega \in [R=x]} R(\omega) \Pr[\omega] \\
 &= \sum_{x \in \text{range}(R)} \sum_{\omega \in [R=x]} x \Pr[\omega] && \text{(def of the event } [R = x]) \\
 &= \sum_{x \in \text{range}(R)} x \left( \sum_{\omega \in [R=x]} \Pr[\omega] \right) && \text{(distributing } x \text{ over the inner sum)} \\
 &= \sum_{x \in \text{range}(R)} x \cdot \Pr[R = x]. && \text{(def of } \Pr[R = x])
 \end{aligned}$$

The first equality follows because the events  $[R = x]$  for  $x \in \text{range}(R)$  partition the sample space  $\mathcal{S}$ , so summing over the outcomes in  $[R = x]$  for  $x \in \text{range}(R)$  is the same as summing over  $\mathcal{S}$ . ■

In general, Equation 18.2 is more useful than Equation 18.1 for calculating expected values and has the advantage that it does not depend on the sample space, but only on the density function of the random variable. It is especially useful when the range of the random variable is  $\mathbb{N}$ , as we will see from the following corollary.

**Corollary 18.1.5.** *If the range of a random variable  $R$  is  $\mathbb{N}$ , then*

$$\text{Ex}[R] = \sum_{i=1}^{\infty} i \Pr[R = i] = \sum_{i=0}^{\infty} \Pr[R > i].$$

*Proof.* The first equality follows directly from Theorem 18.1.4 and the fact that  $\text{range}(R) = \mathbb{N}$ . The second equality is derived by adding the following equations:

$$\begin{array}{rcl}
 \Pr[R > 0] & = & \Pr[R = 1] + \Pr[R = 2] + \Pr[R = 3] + \cdots \\
 \Pr[R > 1] & = & \Pr[R = 2] + \Pr[R = 3] + \cdots \\
 \Pr[R > 2] & = & \Pr[R = 3] + \cdots \\
 & & \vdots \\
 \hline
 \sum_{i=0}^{\infty} \Pr[R > i] & = & 1 \cdot \Pr[R = 1] + 2 \cdot \Pr[R = 2] + 3 \cdot \Pr[R = 3] + \cdots \\
 & = & \sum_{i=1}^{\infty} i \Pr[R = i]. \quad \blacksquare
 \end{array}$$

### 18.1.4 Mean Time to Failure

The mean time to failure is a critical parameter in the design of most any system. For example, suppose that a computer program crashes at the end of each hour of use with probability  $p$ , if it has not crashed already. What is the expected time until the program crashes?

If we let  $C$  be the number of hours until the crash, then the answer to our problem is  $\text{Ex}[C]$ .  $C$  is a random variable with values in  $\mathbb{N}$  and so we can use Corollary 18.1.5 to determine that

$$\text{Ex}[C] = \sum_{i=0}^{\infty} \text{Pr}[C > i]. \quad (18.3)$$

$\text{Pr}[C > i]$  is easy to evaluate: a crash happens later than the  $i$ th hour iff the system did not crash during the first  $i$  hours, which happens with probability  $(1 - p)^i$ . Plugging this into Equation 18.3 gives:

$$\begin{aligned} \text{Ex}[C] &= \sum_{i=0}^{\infty} (1 - p)^i \\ &= \frac{1}{1 - (1 - p)} && \text{(sum of geometric series)} \\ &= \frac{1}{p}. \end{aligned} \quad (18.4)$$

For example, if there is a 1% chance that the program crashes at the end of each hour, then the expected time until the program crashes is  $1/0.01 = 100$  hours.

The general principle here is well-worth remembering:

*If a system fails at each time step with probability  $p$ , then the expected number of steps up to (and including) the first failure is  $1/p$ .*

### Making Babies

As a related example, suppose a couple really wants to have a baby girl. For simplicity, assume that there is a 50% chance that each child they have is a girl, and that the genders of their children are mutually independent. If the couple insists on having children until they get a girl, then how many baby boys should they expect first?

The question, “How many hours until the program crashes?” is mathematically the same as the question, “How many children must the couple have until they get a girl?” In this case, a crash corresponds to having a girl, so we should set

### 18.1. Definitions and Examples

$p = 1/2$ . By the preceding analysis, the couple should expect a baby girl after having  $1/p = 2$  children. Since the last of these will be the girl, they should expect just one boy.

#### 18.1.5 Dealing with Infinity

The analysis of the mean time to failure was easy enough. But if you think about it further, you might start to wonder about the case when the computer program *never* fails. For example, what if the program runs forever? How do we handle outcomes with an infinite value?

These are good questions and we wonder about them too. Indeed, mathematicians have gone to a lot of work to reason about sample spaces with an infinite number of outcomes or outcomes with infinite value.

To keep matters simple in this text, we will follow the common convention of ignoring the contribution of outcomes that have probability zero when computing expected values. This means that we can safely ignore the “never-fail” outcome, because it has probability

$$\lim_{n \rightarrow \infty} (1 - p)^n = 0.$$

In general, when we are computing expectations for infinite sample spaces, we will generally focus our attention on a subset of outcomes that occur with collective probability one. For the most part, this will allow us to ignore the “infinite” outcomes because they will typically happen with probability zero.<sup>2</sup>

This assumption does *not* mean that the expected value of a random variable is always finite, however. Indeed, there are many examples where the expected value is infinite. And where infinity raises its ugly head, trouble is sure to follow. Let’s see an example.

#### 18.1.6 Pitfall: Computing Expectations by Sampling

Suppose that you are trying to estimate a parameter such as the average delay across a communication channel. So you set up an experiment to measure how long it takes to send a test packet from one end to the other and you run the experiment 100 times.

You record the latency, rounded to the nearest millisecond, for each of the hundred experiments, and then compute the average of the 100 measurements. Suppose that this average is 8.3 ms.

Because you are careful, you repeat the entire process twice more and get averages of 7.8 ms and 7.9 ms. You conclude that the average latency across the channel

---

<sup>2</sup>If this still bothers you, you might consider taking a course on measure theory.

Chapter 18 Expectation

is

$$\frac{7.8 + 7.9 + 8.3}{3} = 8 \text{ ms.}$$

You might be right but you might also be horribly wrong. In fact, the expected latency might well be *infinite*. Here’s how.

Let  $D$  be a random variable that denotes the time it takes for the packet to cross the channel. Suppose that

$$\Pr[D = i] = \begin{cases} 0 & \text{for } i = 0 \\ \frac{1}{i} - \frac{1}{i+1} & \text{for } i \in \mathbb{N}^+. \end{cases} \quad (18.5)$$

It is easy to check that

$$\sum_{i=0}^{\infty} \Pr[D = i] = \left(1 - \frac{1}{2}\right) + \left(\frac{1}{2} - \frac{1}{3}\right) + \left(\frac{1}{3} - \frac{1}{4}\right) + \dots = 1$$

and so  $D$  is, in fact, a random variable.

From Equation 18.5, we might expect that  $D$  is likely to be small. Indeed,  $D = 1$  with probability  $1/2$ ,  $D = 2$  with probability  $1/6$ , and so forth. So if we took 100 samples of  $D$ , about 50 would be 1 ms, about 16 would be 2 ms, and very few would be large. In summary, it might well be the case that the average of the 100 measurements would be under 10 ms, just as in our example.

This sort of reasoning and the calculation of expected values by averaging experimental values is very common in practice. It can easily lead to incorrect conclusions, however. For example, using Corollary 18.1.5, we can quickly (and accurately) determine that

$$\begin{aligned} \text{Ex}[D] &= \sum_{i=1}^{\infty} i \Pr[D = i] \\ &= \sum_{i=1}^{\infty} i \left( \frac{1}{i} - \frac{1}{i+1} \right) \\ &= \sum_{i=1}^{\infty} i \left( \frac{1}{i(i+1)} \right) \\ &= \sum_{i=1}^{\infty} \left( \frac{1}{i+1} \right) \\ &= \infty. \end{aligned}$$

Uh-oh! The expected time to cross the communication channel is *infinite*! This result is a far cry from the 10 ms that we calculated. What went wrong?

18.1. Definitions and Examples

It is true that most of the time, the value of  $D$  will be small. But sometimes  $D$  will be very large and this happens with sufficient probability that the expected value of  $D$  is unbounded. In fact, if you keep repeating the experiment, you are likely to see some outcomes and averages that are much larger than 10 ms. In practice, such “outliers” are sometimes discarded, which masks the true behavior of  $D$ .

In general, the best way to compute an expected value in practice is to first use the experimental data to figure out the distribution as best you can, and then to use Theorem 18.1.4 or Corollary 18.1.5 to compute its expectation. This method will help you identify cases where the expectation is infinite, and will generally be more accurate than a simple averaging of the data.

**18.1.7 Conditional Expectation**

Just like event probabilities, expectations can be conditioned on some event. Given a random variable  $R$ , the expected value of  $R$  conditioned on an event  $A$  is the (probability-weighted) average value of  $R$  over outcomes in  $A$ . More formally:

**Definition 18.1.6.** The *conditional expectation*  $\text{Ex}[R \mid A]$  of a random variable  $R$  given event  $A$  is:

$$\text{Ex}[R \mid A] ::= \sum_{r \in \text{range}(R)} r \cdot \Pr[R = r \mid A]. \tag{18.6}$$

For example, we can compute the expected value of a roll of a fair die, *given*, for example, that the number rolled is at least 4. We do this by letting  $R$  be the outcome of a roll of the die. Then by equation (18.6),

$$\text{Ex}[R \mid R \geq 4] = \sum_{i=1}^6 i \cdot \Pr[R = i \mid R \geq 4] = 1 \cdot 0 + 2 \cdot 0 + 3 \cdot 0 + 4 \cdot \frac{1}{3} + 5 \cdot \frac{1}{3} + 6 \cdot \frac{1}{3} = 5.$$

As another example, consider the channel latency problem from Section 18.1.6. The expected latency for this problem was infinite. But what if we look at the

Chapter 18 Expectation

expected latency conditioned on the latency not exceeding  $n$ . Then

$$\begin{aligned}
 \text{Ex}[D] &= \sum_{i=1}^{\infty} i \Pr[D = i \mid D \leq n] \\
 &= \sum_{i=1}^{\infty} i \frac{\Pr[D = i \wedge D \leq n]}{\Pr[D \leq n]} \\
 &= \sum_{i=1}^n \frac{i \Pr[D = i]}{\Pr[D \leq n]} \\
 &= \frac{1}{\Pr[D \leq n]} \sum_{i=1}^n i \left( \frac{1}{i(i+1)} \right) \\
 &= \frac{1}{\Pr[D \leq n]} \sum_{i=1}^n \frac{1}{i+1} \\
 &= \frac{1}{\Pr[D \leq n]} (H_{n+1} - 1),
 \end{aligned}$$

where  $H_{n+1}$  is the  $(n + 1)$ st Harmonic number

$$H_{n+1} = \ln(n + 1) + \gamma + \frac{1}{2n} + \frac{1}{12n^2} + \frac{\epsilon(n)}{120n^4}$$

and  $0 \leq \epsilon(n) \leq 1$ . The second equality follows from the definition of conditional expectation, the third equality follows from the fact that  $\Pr[D = i \wedge D \leq n] = 0$  for  $i > n$ , and the fourth equality follows from the definition of  $D$  in Equation 18.5.

To compute  $\Pr[D \leq n]$ , we observe that

$$\begin{aligned}
 \Pr[D \leq n] &= 1 - \Pr[D > n] \\
 &= 1 - \sum_{i=n+1}^{\infty} \left( \frac{1}{i} - \frac{1}{i+1} \right) \\
 &= 1 - \left[ \left( \frac{1}{n+1} - \frac{1}{n+2} \right) + \left( \frac{1}{n+2} - \frac{1}{n+3} \right) \right. \\
 &\quad \left. + \left( \frac{1}{n+3} - \frac{1}{n+4} \right) + \dots \right] \\
 &= 1 - \frac{1}{n+1} \\
 &= \frac{n}{n+1}.
 \end{aligned}$$

18.1. Definitions and Examples

Hence,

$$\text{Ex}[D] = \frac{n+1}{n}(H_{n+1} - 1). \quad (18.7)$$

For  $n = 1000$ , this is about 6.5. This explains why the expected value of  $D$  appears to be finite when you try to evaluate it experimentally. If you compute 100 samples of  $D$ , it is likely that all of them will be at most 1000 ms. If you condition on not having any outcomes greater than 1000 ms, then the conditional expected value will be about 6.5 ms, which would be a commonly observed result in practice. Yet we know that  $\text{Ex}[D]$  is infinite. For this reason, expectations computed in practice are often really just conditional expectations where the condition is that rare “outlier” sample points are eliminated from the analysis.

### 18.1.8 The Law of Total Expectation

Another useful feature of conditional expectation is that it lets us divide complicated expectation calculations into simpler cases. We can then find the desired expectation by calculating the conditional expectation in each simple case and averaging them, weighing each case by its probability.

For example, suppose that 49.8% of the people in the world are male and the rest female—which is more or less true. Also suppose the expected height of a randomly chosen male is 5' 11", while the expected height of a randomly chosen female is 5' 5". What is the expected height of a randomly chosen individual? We can calculate this by averaging the heights of men and women. Namely, let  $H$  be the height (in feet) of a randomly chosen person, and let  $M$  be the event that the person is male and  $F$  the event that the person is female. Then

$$\begin{aligned} \text{Ex}[H] &= \text{Ex}[H \mid M] \Pr[M] + \text{Ex}[H \mid F] \Pr[F] \\ &= (5 + 11/12) \cdot 0.498 + (5 + 5/12) \cdot 0.502 \\ &= 5.665 \end{aligned}$$

which is a little less than 5' 8".

This method is justified by the Law of *Total Expectation*.

**Theorem 18.1.7** (Law of Total Expectation). *Let  $R$  be a random variable on a sample space  $\mathcal{S}$  and suppose that  $A_1, A_2, \dots$ , is a partition of  $\mathcal{S}$ . Then*

$$\text{Ex}[R] = \sum_i \text{Ex}[R \mid A_i] \Pr[A_i].$$

Chapter 18 Expectation

*Proof.*

$$\begin{aligned}
 \text{Ex}[R] &= \sum_{r \in \text{range}(R)} r \cdot \Pr[R = r] && \text{(Equation 18.2)} \\
 &= \sum_r r \cdot \sum_i \Pr[R = r \mid A_i] \Pr[A_i] && \text{(Law of Total Probability)} \\
 &= \sum_r \sum_i r \cdot \Pr[R = r \mid A_i] \Pr[A_i] && \text{(distribute constant } r) \\
 &= \sum_i \sum_r r \cdot \Pr[R = r \mid A_i] \Pr[A_i] && \text{(exchange order of summation)} \\
 &= \sum_i \Pr[A_i] \sum_r r \cdot \Pr[R = r \mid A_i] && \text{(factor constant } \Pr[A_i]) \\
 &= \sum_i \Pr[A_i] \text{Ex}[R \mid A_i]. && \text{(Def 18.1.6 of cond. expectation)}
 \end{aligned}$$

■

As a more interesting application of the Law of Total Expectation, let’s take another look at the mean time to failure of a system that fails with probability  $p$  at each step. We’ll define  $A$  to be the event that the system fails on the first step and  $\bar{A}$  to be the complementary event (namely, that the system does not fail on the first step). Then the mean time to failure  $\text{Ex}[C]$  is

$$\text{Ex}[C] = \text{Ex}[C \mid A] \Pr[A] + \text{Ex}[C \mid \bar{A}] \Pr[\bar{A}]. \quad (18.8)$$

Since  $A$  is the condition that the system crashes on the first step, we know that

$$\text{Ex}[C \mid A] = 1. \quad (18.9)$$

Since  $\bar{A}$  is the condition that the system does *not* crash on the first step, conditioning on  $\bar{A}$  is equivalent to taking a first step without failure and then starting over without conditioning. Hence,

$$\text{Ex}[C \mid \bar{A}] = 1 + \text{Ex}[C]. \quad (18.10)$$

Plugging Equations 18.9 and 18.10 into Equation 18.8, we find that

$$\begin{aligned}
 \text{Ex}[C] &= 1 \cdot p + (1 + \text{Ex}[C])(1 - p) \\
 &= p + 1 - p + (1 - p) \text{Ex}[C] \\
 &= 1 + (1 - p) \text{Ex}[C].
 \end{aligned}$$

18.2. *Expected Returns in Gambling Games*

Rearranging terms, we find that

$$1 = \text{Ex}[C] - (1 - p) \text{Ex}[C] = p \text{Ex}[C],$$

and thus that

$$\text{Ex}[C] = \frac{1}{p},$$

as expected.

We will use this sort of analysis extensively in Chapter 20 when we examine the expected behavior of random walks.

**18.1.9 Expectations of Functions**

Expectations can also be defined for functions of random variables.

**Definition 18.1.8.** Let  $R : \mathcal{S} \rightarrow V$  be a random variable and  $f : V \rightarrow \mathbb{R}$  be a total function on the range of  $R$ . Then

$$\text{Ex}[f(R)] = \sum_{w \in \mathcal{S}} f(R(w)) \text{Pr}[w]. \tag{18.11}$$

Equivalently,

$$\text{Ex}[f(R)] = \sum_{r \in \text{range}(R)} f(r) \text{Pr}[R = r]. \tag{18.12}$$

For example, suppose that  $R$  is the value obtained by rolling a fair 6-sided die. Then

$$\text{Ex} \left[ \frac{1}{R} \right] = \frac{1}{1} \cdot \frac{1}{6} + \frac{1}{2} \cdot \frac{1}{6} + \frac{1}{3} \cdot \frac{1}{6} + \frac{1}{4} \cdot \frac{1}{6} + \frac{1}{5} \cdot \frac{1}{6} + \frac{1}{6} \cdot \frac{1}{6} = \frac{49}{120}.$$

**18.2 Expected Returns in Gambling Games**

Some of the most interesting examples of expectation can be explained in terms of gambling games. For straightforward games where you win  $\$A$  with probability  $p$  and you lose  $\$B$  with probability  $1 - p$ , it is easy to compute your *expected return* or *winnings*. It is simply

$$pA - (1 - p)B.$$

For example, if you are flipping a fair coin and you win  $\$1$  for heads and you lose  $\$1$  for tails, then your expected winnings are

$$\frac{1}{2} \cdot 1 - \left(1 - \frac{1}{2}\right) \cdot 1 = 0.$$

Chapter 18 Expectation

In such cases, the game is said to be *fair* since your expected return is zero.

Some gambling games are more complicated and thus more interesting. For example, consider the following game where the winners split a pot. This sort of game is representative of many poker games, betting pools, and lotteries.

### 18.2.1 Splitting the Pot

After your last encounter with biker dude, one thing lead to another and you have dropped out of school and become a Hell’s Angel. It’s late on a Friday night and, feeling nostalgic for the old days, you drop by your old hangout, where you encounter two of your former TAs, Eric and Nick. Eric and Nick propose that you join them in a simple wager. Each player will put \$2 on the bar and secretly write “heads” or “tails” on their napkin. Then one player will flip a fair coin. The \$6 on the bar will then be divided equally among the players who correctly predicted the outcome of the coin toss.

After your life-altering encounter with strange dice, you are more than a little skeptical. So Eric and Nick agree to let you be the one to flip the coin. This certainly seems fair. How can you lose?

But you have learned your lesson and so before agreeing, you go through the four-step method and write out the tree diagram to compute your expected return. The tree diagram is shown in Figure 18.1.

The “payoff” values in Figure 18.1 are computed by dividing the \$6 pot<sup>3</sup> among those players who guessed correctly and then subtracting the \$2 that you put into the pot at the beginning. For example, if all three players guessed correctly, then you payoff is \$0, since you just get back your \$2 wager. If you and Nick guess correctly and Eric guessed wrong, then your payoff is

$$\frac{6}{2} - 2 = 1.$$

In the case that everyone is wrong, you all agree to split the pot and so, again, your payoff is zero.

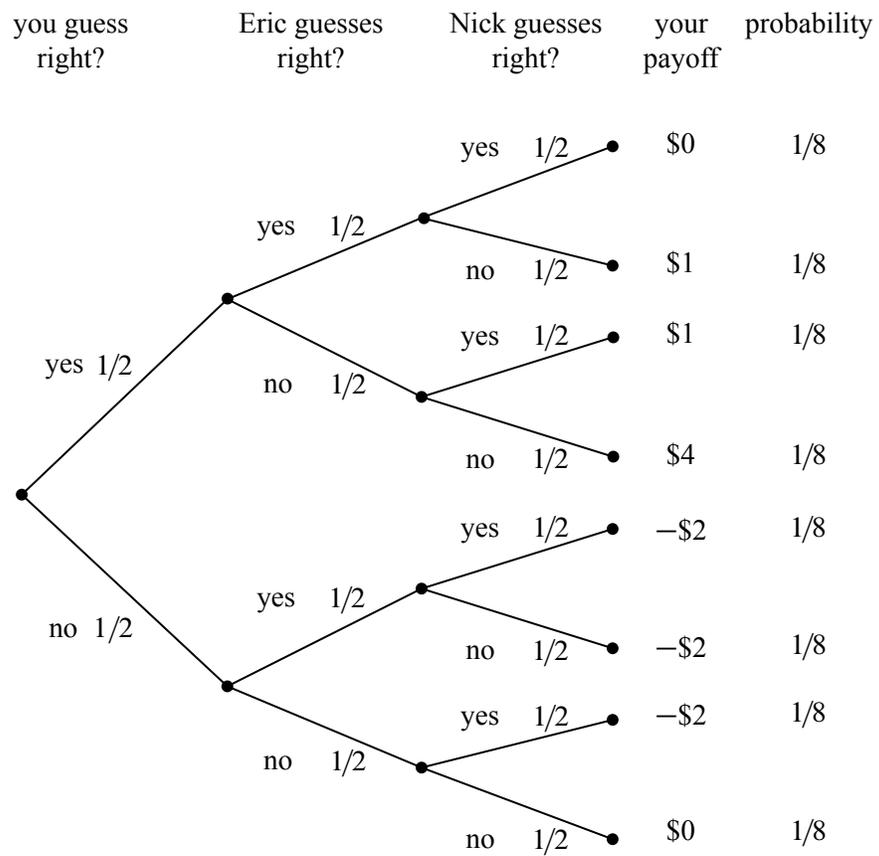
To compute your expected return, you use Equation 18.1 in the definition of expected value. This yields

$$\begin{aligned} \text{Ex}[\text{payoff}] &= 0 \cdot \frac{1}{8} + 1 \cdot \frac{1}{8} + 1 \cdot \frac{1}{8} + 4 \cdot \frac{1}{8} \\ &\quad + (-2) \cdot \frac{1}{8} + (-2) \cdot \frac{1}{8} + (-2) \cdot \frac{1}{8} + 0 \cdot \frac{1}{8} \\ &= 0. \end{aligned}$$

---

<sup>3</sup>The money invested in a wager is commonly referred to as the *pot*.

18.2. Expected Returns in Gambling Games



**Figure 18.1** The tree diagram for the game where three players each wager \$2 and then guess the outcome of a fair coin toss. The winners split the pot.

Chapter 18 *Expectation*

This confirms that the game is fair. So, for old time’s sake, you break your solemn vow to never ever engage in strange gambling games.

**18.2.2 The Impact of Collusion**

Needless to say, things are not turning out well for you. The more times you play the game, the more money you seem to be losing. After 1000 wagers, you have lost over \$500. As Nick and Eric are consoling you on your “bad luck,” you do a back-of-the-napkin calculation using the bounds on the tails of the binomial distribution from Section 17.5 that suggests that the probability of losing \$500 in 1000 wagers is less than the probability of a Vietnamese Monk waltzing in and handing you one of those golden disks. How can this be?

It is possible that you are truly very very unlucky. But it is more likely that something is wrong with the tree diagram in Figure 18.1 and that “something” just might have something to do with the possibility that Nick and Eric are colluding against you.

To be sure, Nick and Eric can only guess the outcome of the coin toss with probability 1/2, but what if Nick and Eric always guess differently? In other words, what if Nick always guesses “tails” when Eric guesses “heads,” and vice-versa? This would result in a slightly different tree diagram, as shown in Figure 18.2.

The payoffs for each outcome are the same in Figures 18.1 and 18.2, but the probabilities of the outcomes are different. For example, it is no longer possible for all three players to guess correctly, since Nick and Eric are always guessing differently. More importantly, the outcome where your payoff is \$4 is also no longer possible. Since Nick and Eric are always guessing differently, one of them will always get a share of the pot. As you might imagine, this is not good for you!

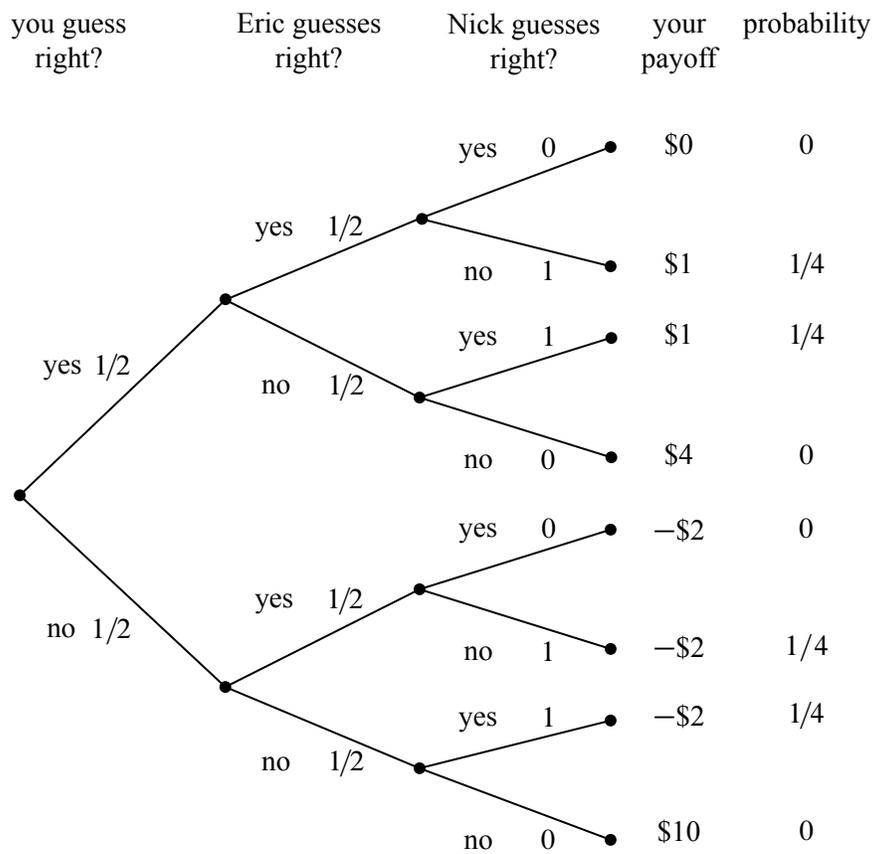
When we use Equation 18.1 to compute your expected return in the collusion scenario, we find that

$$\begin{aligned} \text{Ex}[\text{payoff}] &= 0 \cdot 0 + 1 \cdot \frac{1}{4} + 1 \cdot \frac{1}{4} + 4 \cdot 0 \\ &\quad + (-2) \cdot 0 + (-2) \cdot \frac{1}{4} + (-2) \cdot \frac{1}{4} + 0 \cdot 0 \\ &= -\frac{1}{2}. \end{aligned}$$

This is very bad indeed. By colluding, Nick and Eric have made it so that you expect to lose \$.50 every time you play. No wonder you lost \$500 over the course of 1000 wagers.

Maybe it would be a good idea to go back to school—your Hell’s Angels buds may not be too happy that you just lost their \$500.

18.2. Expected Returns in Gambling Games



**Figure 18.2** The revised tree diagram reflecting the scenario where Nick always guesses the opposite of Eric.

### 18.2.3 How to Win the Lottery

Similar opportunities to “collude” arise in many betting games. For example, consider the typical weekly football betting pool, where each participant wagers \$10 and the participants that pick the most games correctly split a large pot. The pool seems fair if you think of it as in Figure 18.1. But, in fact, if two or more players collude by guessing differently, they can get an “unfair” advantage at your expense!

In some cases, the collusion is inadvertent and you can profit from it. For example, many years ago, a former MIT Professor of Mathematics named Herman Chernoff figured out a way to make money by playing the state lottery. This was surprising since state lotteries typically have very poor expected returns. That’s because the state usually takes a large share of the wagers before distributing the rest of the pot among the winners. Hence, anyone who buys a lottery ticket is expected to *lose* money. So how did Chernoff find a way to make money? It turned out to be easy!

In a typical state lottery,

- all players pay \$1 to play and select 4 numbers from 1 to 36,
- the state draws 4 numbers from 1 to 36 uniformly at random,
- the state divides 1/2 of the money collected among the people who guessed correctly and spends the other half redecorating the governor’s residence.

This is a lot like the game you played with Nick and Eric, except that there are more players and more choices. Chernoff discovered that a small set of numbers was selected by a large fraction of the population. Apparently many people think the same way; they pick the same numbers not on purpose as in the previous game with Nick and Eric, but based on Manny’s batting average or today’s date.

It was as if the players were colluding to lose! If any one of them guessed correctly, then they’d have to split the pot with many other players. By selecting numbers uniformly at random, Chernoff was unlikely to get one of these favored sequences. So if he won, he’d likely get the whole pot! By analyzing actual state lottery data, he determined that he could win an average of 7 cents on the dollar. In other words, his expected return was not  $-\$.50$  as you might think, but  $+\$.07$ .<sup>4</sup>

Inadvertent collusion often arises in betting pools and is a phenomenon that you can take advantage of. For example, suppose you enter a Super Bowl betting pool where the goal is to get closest to the total number of points scored in the game. Also suppose that the average Super Bowl has a total of 30 point scored and that

---

<sup>4</sup>Most lotteries now offer randomized tickets to help smooth out the distribution of selected sequences.

18.3. Expectations of Sums

everyone knows this. Then most people will guess around 30 points. Where should you guess? Well, you should guess just outside of this range because you get to cover a lot more ground and you don’t share the pot if you win. Of course, if you are in a pool with math students and they all know this strategy, then maybe you should guess 30 points after all.

## 18.3 Expectations of Sums

### 18.3.1 Linearity of Expectation

Expected values obey a simple, very helpful rule called *Linearity of Expectation*. Its simplest form says that the expected value of a sum of random variables is the sum of the expected values of the variables.

**Theorem 18.3.1.** *For any random variables  $R_1$  and  $R_2$ ,*

$$\text{Ex}[R_1 + R_2] = \text{Ex}[R_1] + \text{Ex}[R_2].$$

*Proof.* Let  $T ::= R_1 + R_2$ . The proof follows straightforwardly by rearranging terms in Equation (18.1):

$$\begin{aligned} \text{Ex}[T] &= \sum_{\omega \in \mathcal{S}} T(\omega) \cdot \text{Pr}[\omega] && \text{(Definition 18.1.1)} \\ &= \sum_{\omega \in \mathcal{S}} (R_1(\omega) + R_2(\omega)) \cdot \text{Pr}[\omega] && \text{(definition of } T) \\ &= \sum_{\omega \in \mathcal{S}} R_1(\omega) \text{Pr}[\omega] + \sum_{\omega \in \mathcal{S}} R_2(\omega) \text{Pr}[\omega] && \text{(rearranging terms)} \\ &= \text{Ex}[R_1] + \text{Ex}[R_2]. && \text{(Definition 18.1.1)} \quad \blacksquare \end{aligned}$$

A small extension of this proof, which we leave to the reader, implies

**Theorem 18.3.2.** *For random variables  $R_1, R_2$  and constants  $a_1, a_2 \in \mathbb{R}$ ,*

$$\text{Ex}[a_1 R_1 + a_2 R_2] = a_1 \text{Ex}[R_1] + a_2 \text{Ex}[R_2].$$

In other words, expectation is a linear function. A routine induction extends the result to more than two variables:

**Corollary 18.3.3** (Linearity of Expectation). *For any random variables  $R_1, \dots, R_k$  and constants  $a_1, \dots, a_k \in \mathbb{R}$ ,*

$$\text{Ex}\left[\sum_{i=1}^k a_i R_i\right] = \sum_{i=1}^k a_i \text{Ex}[R_i].$$

Chapter 18 Expectation

The great thing about linearity of expectation is that *no independence is required*. This is really useful, because dealing with independence is a pain, and we often need to work with random variables that are not known to be independent.

As an example, let’s compute the expected value of the sum of two fair dice. Let the random variable  $R_1$  be the number on the first die, and let  $R_2$  be the number on the second die. We observed earlier that the expected value of one die is 3.5. We can find the expected value of the sum using linearity of expectation:

$$\text{Ex}[R_1 + R_2] = \text{Ex}[R_1] + \text{Ex}[R_2] = 3.5 + 3.5 = 7.$$

Notice that we did *not* have to assume that the two dice were independent. The expected sum of two dice is 7, even if they are glued together (provided each individual die remains fair after the gluing). Proving that this expected sum is 7 with a tree diagram would be a bother: there are 36 cases. And if we did not assume that the dice were independent, the job would be really tough!

### 18.3.2 Sums of Indicator Random Variables

Linearity of expectation is especially useful when you have a sum of indicator random variables. As an example, suppose there is a dinner party where  $n$  men check their hats. The hats are mixed up during dinner, so that afterward each man receives a random hat. In particular, each man gets his own hat with probability  $1/n$ . What is the expected number of men who get their own hat?

Letting  $G$  be the number of men that get their own hat, we want to find the expectation of  $G$ . But all we know about  $G$  is that the probability that a man gets his own hat back is  $1/n$ . There are many different probability distributions of hat permutations with this property, so we don’t know enough about the distribution of  $G$  to calculate its expectation directly. But linearity of expectation makes the problem really easy.

The trick<sup>5</sup> is to express  $G$  as a sum of indicator variables. In particular, let  $G_i$  be an indicator for the event that the  $i$ th man gets his own hat. That is,  $G_i = 1$  if the  $i$ th man gets his own hat, and  $G_i = 0$  otherwise. The number of men that get their own hat is then the sum of these indicator random variables:

$$G = G_1 + G_2 + \cdots + G_n. \tag{18.13}$$

These indicator variables are *not* mutually independent. For example, if  $n - 1$  men all get their own hats, then the last man is certain to receive his own hat. But, since we plan to use linearity of expectation, we don’t have worry about independence!

---

<sup>5</sup>We are going to use this trick a lot so it is important to understand it.

18.3. Expectations of Sums

Since  $G_i$  is an indicator random variable, we know from Lemma 18.1.3 that

$$\text{Ex}[G_i] = \Pr[G_i = 1] = 1/n. \quad (18.14)$$

By Linearity of Expectation and Equation 18.13, this means that

$$\begin{aligned} \text{Ex}[G] &= \text{Ex}[G_1 + G_2 + \cdots + G_n] \\ &= \text{Ex}[G_1] + \text{Ex}[G_2] + \cdots + \text{Ex}[G_n] \\ &= \overbrace{\frac{1}{n} + \frac{1}{n} + \cdots + \frac{1}{n}}^n \\ &= 1. \end{aligned}$$

So even though we don't know much about how hats are scrambled, we've figured out that on average, just one man gets his own hat back!

More generally, Linearity of Expectation provides a very good method for computing the expected number of events that will happen.

**Theorem 18.3.4.** *Given any collection of  $n$  events  $A_1, A_2, \dots, A_n \subseteq \mathcal{S}$ , the expected number of events that will occur is*

$$\sum_{i=1}^n \Pr[A_i].$$

For example,  $A_i$  could be the event that the  $i$ th man gets the right hat back. But in general, it could be any subset of the sample space, and we are asking for the expected number of events that will contain a random sample point.

*Proof.* Define  $R_i$  to be the indicator random variable for  $A_i$ , where  $R_i(w) = 1$  if  $w \in A_i$  and  $R_i(w) = 0$  if  $w \notin A_i$ . Let  $R = R_1 + R_2 + \cdots + R_n$ . Then

$$\begin{aligned} \text{Ex}[R] &= \sum_{i=1}^n \text{Ex}[R_i] && \text{(by Linearity of Expectation)} \\ &= \sum_{i=1}^n \Pr[R_i = 1] && \text{(by Lemma 18.1.3)} \\ &= \sum_{i=1}^n \sum_{w \in A_i} \Pr[w] && \text{(definition of indicator variable)} \\ &= \sum_{i=1}^n \Pr[A_i]. \quad \blacksquare \end{aligned}$$

Chapter 18 Expectation

So whenever you are asked for the expected number of events that occur, all you have to do is sum the probabilities that each event occurs. Independence is not needed.

### 18.3.3 Expectation of a Binomial Distribution

Suppose that we independently flip  $n$  biased coins, each with probability  $p$  of coming up heads. What is the expected number of heads?

Let  $J$  be the random variable denoting the number of heads. Then  $J$  has a binomial distribution with parameters  $n$ ,  $p$ , and

$$\Pr[J = k] = \binom{n}{k} k^p (n - k)^{1-p}.$$

Applying Equation 18.2, this means that

$$\begin{aligned} \text{Ex}[J] &= \sum_{k=0}^n k \Pr[J = k] \\ &= \sum_{k=0}^n k \binom{n}{k} k^p (n - k)^{1-p}. \end{aligned} \tag{18.15}$$

Ouch! This is one nasty looking sum. Let’s try another approach.

Since we have just learned about linearity of expectation for sums of indicator random variables, maybe Theorem 18.3.4 will be helpful. But how do we express  $J$  as a sum of indicator random variables? It turns out to be easy. Let  $J_i$  be the indicator random variable for the  $i$ th coin. In particular, define

$$J_i = \begin{cases} 1 & \text{if the } i\text{th coin is heads} \\ 0 & \text{if the } i\text{th coin is tails.} \end{cases}$$

Then the number of heads is simply

$$J = J_1 + J_2 + \cdots + J_n.$$

By Theorem 18.3.4,

$$\begin{aligned} \text{Ex}[J] &= \sum_{i=1}^n \Pr[J_i] \\ &= np. \end{aligned} \tag{18.16}$$

18.3. Expectations of Sums

That really was easy. If we flip  $n$  mutually independent coins, we expect to get  $pn$  heads. Hence the expected value of a binomial distribution with parameters  $n$  and  $p$  is simply  $pn$ .

But what if the coins are not mutually independent? It doesn't matter—the answer is still  $pn$  because Linearity of Expectation and Theorem 18.3.4 do not assume any independence.

If you are not yet convinced that Linearity of Expectation and Theorem 18.3.4 are powerful tools, consider this: without even trying, we have used them to prove a very complicated identity, namely<sup>6</sup>

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

If you are still not convinced, then take a look at the next problem.

**18.3.4 The Coupon Collector Problem**

Every time we purchase a kid's meal at Taco Bell, we are graciously presented with a miniature “Racin' Rocket” car together with a launching device which enables us to project our new vehicle across any tabletop or smooth floor at high velocity. Truly, our delight knows no bounds.

There are  $n$  different types of Racin' Rocket cars (blue, green, red, gray, etc.). The type of car awarded to us each day by the kind woman at the Taco Bell register appears to be selected uniformly and independently at random. What is the expected number of kid's meals that we must purchase in order to acquire at least one of each type of Racin' Rocket car?

The same mathematical question shows up in many guises: for example, what is the expected number of people you must poll in order to find at least one person with each possible birthday? Here, instead of collecting Racin' Rocket cars, you're collecting birthdays. The general question is commonly called the *coupon collector problem* after yet another interpretation.

A clever application of linearity of expectation leads to a simple solution to the coupon collector problem. Suppose there are five different types of Racin' Rocket cars, and we receive this sequence:

blue green green red blue orange blue orange gray.

Let's partition the sequence into 5 segments:

$\underbrace{\text{blue}}_{X_0}$   $\underbrace{\text{green}}_{X_1}$   $\underbrace{\text{green red}}_{X_2}$   $\underbrace{\text{blue orange}}_{X_3}$   $\underbrace{\text{blue orange gray}}_{X_4}$ .

<sup>6</sup>This follows by combining Equations 18.15 and 18.16.

Chapter 18 Expectation

The rule is that a segment ends whenever we get a new kind of car. For example, the middle segment ends when we get a red car for the first time. In this way, we can break the problem of collecting every type of car into stages. Then we can analyze each stage individually and assemble the results using linearity of expectation.

Let’s return to the general case where we’re collecting  $n$  Racin’ Rockets. Let  $X_k$  be the length of the  $k$ th segment. The total number of kid’s meals we must purchase to get all  $n$  Racin’ Rockets is the sum of the lengths of all these segments:

$$T = X_0 + X_1 + \cdots + X_{n-1}$$

Now let’s focus our attention on  $X_k$ , the length of the  $k$ th segment. At the beginning of segment  $k$ , we have  $k$  different types of car, and the segment ends when we acquire a new type. When we own  $k$  types, each kid’s meal contains a type that we already have with probability  $k/n$ . Therefore, each meal contains a new type of car with probability  $1 - k/n = (n - k)/n$ . Thus, the expected number of meals until we get a new kind of car is  $n/(n - k)$  by the “mean time to failure” formula in Equation 18.4. This means that

$$\text{Ex}[X_k] = \frac{n}{n - k}.$$

Linearity of expectation, together with this observation, solves the coupon collector problem:

$$\begin{aligned} \text{Ex}[T] &= \text{Ex}[X_0 + X_1 + \cdots + X_{n-1}] \\ &= \text{Ex}[X_0] + \text{Ex}[X_1] + \cdots + \text{Ex}[X_{n-1}] \\ &= \frac{n}{n - 0} + \frac{n}{n - 1} + \cdots + \frac{n}{3} + \frac{n}{2} + \frac{n}{1} \\ &= n \left( \frac{1}{n} + \frac{1}{n - 1} + \cdots + \frac{1}{3} + \frac{1}{2} + \frac{1}{1} \right) \\ &= n \left( \frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n - 1} + \frac{1}{n} \right) \\ &= nH_n \end{aligned} \tag{18.17}$$

$$\sim n \ln n. \tag{18.18}$$

Wow! It’s those Harmonic Numbers again!

We can use Equation 18.18 to answer some concrete questions. For example, the expected number of die rolls required to see every number from 1 to 6 is:

$$6H_6 = 14.7 \dots$$

18.3. Expectations of Sums

And the expected number of people you must poll to find at least one person with each possible birthday is:

$$365H_{365} = 2364.6\dots$$

**18.3.5 Infinite Sums**

Linearity of expectation also works for an infinite number of random variables provided that the variables satisfy some stringent absolute convergence criteria.

**Theorem 18.3.5** (Linearity of Expectation). *Let  $R_0, R_1, \dots$ , be random variables such that*

$$\sum_{i=0}^{\infty} \text{Ex}[|R_i|]$$

*converges. Then*

$$\text{Ex} \left[ \sum_{i=0}^{\infty} R_i \right] = \sum_{i=0}^{\infty} \text{Ex}[R_i].$$

*Proof.* Let  $T ::= \sum_{i=0}^{\infty} R_i$ .

We leave it to the reader to verify that, under the given convergence hypothesis, all the sums in the following derivation are absolutely convergent, which justifies rearranging them as follows:

$$\begin{aligned} \sum_{i=0}^{\infty} \text{Ex}[R_i] &= \sum_{i=0}^{\infty} \sum_{s \in \mathcal{S}} R_i(s) \cdot \text{Pr}[s] && \text{(Def. 18.1.1)} \\ &= \sum_{s \in \mathcal{S}} \sum_{i=0}^{\infty} R_i(s) \cdot \text{Pr}[s] && \text{(exchanging order of summation)} \\ &= \sum_{s \in \mathcal{S}} \left[ \sum_{i=0}^{\infty} R_i(s) \right] \cdot \text{Pr}[s] && \text{(factoring out Pr[s])} \\ &= \sum_{s \in \mathcal{S}} T(s) \cdot \text{Pr}[s] && \text{(Def. of } T) \\ &= \text{Ex}[T] && \text{(Def. 18.1.1)} \\ &= \text{Ex} \left[ \sum_{i=0}^{\infty} R_i \right]. && \text{(Def. of } T). \blacksquare \end{aligned}$$

## 18.4 Expectations of Products

While the expectation of a sum is the sum of the expectations, the same is usually not true for products. For example, suppose that we roll a fair 6-sided die and denote the outcome with the random variable  $R$ . Does  $\text{Ex}[R \cdot R] = \text{Ex}[R] \cdot \text{Ex}[R]$ ?

We know that  $\text{Ex}[R] = 3\frac{1}{2}$  and thus  $\text{Ex}[R]^2 = 12\frac{1}{4}$ . Let's compute  $\text{Ex}[R^2]$  to see if we get the same result.

$$\begin{aligned} \text{Ex}[R^2] &= \sum_{w \in \mathcal{S}} R^2(w) \Pr[w] \\ &= \sum_{i=1}^6 i^2 \cdot \Pr[R_i = i] \\ &= \frac{1^2}{6} + \frac{2^2}{6} + \frac{3^2}{6} + \frac{4^2}{6} + \frac{5^2}{6} + \frac{6^2}{6} \\ &= 15 \frac{1}{6} \\ &\neq 12 \frac{1}{4}. \end{aligned}$$

Hence,

$$\text{Ex}[R \cdot R] \neq \text{Ex}[R] \cdot \text{Ex}[R]$$

and so the expectation of a product is not always equal to the product of the expectations.

There is a special case when such a relationship *does* hold however; namely, when the random variables in the product are *independent*.

**Theorem 18.4.1.** *For any two independent random variables  $R_1, R_2$ ,*

$$\text{Ex}[R_1 \cdot R_2] = \text{Ex}[R_1] \cdot \text{Ex}[R_2].$$

*Proof.* The event  $[R_1 \cdot R_2 = r]$  can be split up into events of the form  $[R_1 =$

18.4. Expectations of Products

$r_1$  and  $R_2 = r_2$ ] where  $r_1 \cdot r_2 = r$ . So

$$\begin{aligned}
 & \text{Ex}[R_1 \cdot R_2] \\
 &= \sum_{r \in \text{range}(R_1 \cdot R_2)} r \cdot \Pr[R_1 \cdot R_2 = r] && \text{(Theorem 18.1.4)} \\
 &= \sum_{r_1 \in \text{range}(R_1)} \sum_{r_2 \in \text{range}(R_2)} r_1 r_2 \cdot \Pr[R_1 = r_1 \text{ and } R_2 = r_2] \\
 &= \sum_{r_1 \in \text{range}(R_1)} \sum_{r_2 \in \text{range}(R_2)} r_1 r_2 \cdot \Pr[R_1 = r_1] \cdot \Pr[R_2 = r_2] && \text{(independence of } R_1, R_2) \\
 &= \sum_{r_1 \in \text{range}(R_1)} r_1 \Pr[R_1 = r_1] \left( \sum_{r_2 \in \text{range}(R_2)} r_2 \Pr[R_2 = r_2] \right) && \text{(factor out } r_1 \Pr[R_1 = r_1]) \\
 &= \sum_{r_1 \in \text{range}(R_1)} r_1 \Pr[R_1 = r_1] \cdot \text{Ex}[R_2] && \text{(Theorem 18.1.4)} \\
 &= \text{Ex}[R_2] \left( \sum_{r_1 \in \text{range}(R_1)} r_1 \Pr[R_1 = r_1] \right) && \text{(factor out Ex}[R_2]) \\
 &= \text{Ex}[R_2] \cdot \text{Ex}[R_1]. && \text{(Theorem 18.1.4)}
 \end{aligned}$$

■

For example, let  $R_1$  and  $R_2$  be random variables denoting the result of rolling two independent and fair 6-sided dice. Then

$$\text{Ex}[R_1 \cdot R_2] = \text{Ex}[R_1] \text{Ex}[R_2] = 3\frac{1}{2} \cdot 3\frac{1}{2} = 12\frac{1}{4}.$$

Theorem 18.4.1 extends by induction to a collection of mutually independent random variables.

**Corollary 18.4.2.** *If random variables  $R_1, R_2, \dots, R_k$  are mutually independent, then*

$$\text{Ex} \left[ \prod_{i=1}^k R_i \right] = \prod_{i=1}^k \text{Ex}[R_i].$$

## 18.5 Expectations of Quotients

If  $S$  and  $T$  are random variables, we know from Linearity of Expectation that

$$\text{Ex}[S + T] = \text{Ex}[S] + \text{Ex}[T].$$

If  $S$  and  $T$  are independent, we know from Theorem 18.4.1 that

$$\text{Ex}[ST] = \text{Ex}[S] \text{Ex}[T].$$

Is it also true that

$$\text{Ex}[S/T] = \text{Ex}[S]/\text{Ex}[T]? \tag{18.19}$$

Of course, we have to worry about the situation when  $\text{Ex}[T] = 0$ , but what if we assume that  $T$  is always positive? As we will soon see, Equation 18.19 is usually not true, but let’s see if we can prove it anyway.

**False Claim 18.5.1.** *If  $S$  and  $T$  are independent random variables with  $T > 0$ , then*

$$\text{Ex}[S/T] = \text{Ex}[S]/\text{Ex}[T]. \tag{18.20}$$

*Bogus proof.*

$$\begin{aligned} \text{Ex}\left[\frac{S}{T}\right] &= \text{Ex}\left[S \cdot \frac{1}{T}\right] \\ &= \text{Ex}[S] \cdot \text{Ex}\left[\frac{1}{T}\right] && \text{(independence of } S \text{ and } T) \end{aligned} \tag{18.21}$$

$$= \text{Ex}[S] \cdot \frac{1}{\text{Ex}[T]}. \tag{18.22}$$

$$= \frac{\text{Ex}[S]}{\text{Ex}[T]}. \quad \blacksquare$$

Note that line 18.21 uses the fact that if  $S$  and  $T$  are independent, then so are  $S$  and  $1/T$ . This holds because functions of independent random variables are independent. It is a fact that needs proof, which we will leave to the reader, but it is not the bug. The bug is in line (18.22), which assumes

**False Claim 18.5.2.**

$$\text{Ex}\left[\frac{1}{T}\right] = \frac{1}{\text{Ex}[T]}.$$

18.5. Expectations of Quotients

Benchmark	RISC	CISC	CISC/RISC
E-string search	150	120	0.8
F-bit test	120	180	1.5
Ackerman	150	300	2.0
Rec 2-sort	2800	1400	0.5
Average			1.2

**Table 18.1** Sample program lengths for benchmark problems using RISC and CISC compilers.

Here is a counterexample. Define  $T$  so that

$$\Pr[T = 1] = \frac{1}{2} \quad \text{and} \quad \Pr[T = 2] = \frac{1}{2}.$$

Then

$$\text{Ex}[T] = 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{2} = \frac{3}{2}$$

and

$$\frac{1}{\text{Ex}[T]} = \frac{2}{3}$$

and

$$\text{Ex}\left[\frac{1}{T}\right] = \frac{1}{1} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} = \frac{3}{4} \neq \frac{1}{\text{Ex}[1/T]}.$$

This means that Claim 18.5.1 is also false since we could define  $S = 1$  with probability 1. In fact, both Claims 18.5.1 and 18.5.2 are untrue for most all choices of  $S$  and  $T$ . Unfortunately, the fact that they are false does not keep them from being widely used in practice! Let’s see an example.

**18.5.1 A RISC Paradox**

The data in Table 18.1 is representative of data in a paper by some famous professors. They wanted to show that programs on a RISC processor are generally shorter than programs on a CISC processor. For this purpose, they applied a RISC compiler and then a CISC compiler to some benchmark source programs and made a table of compiled program lengths.

Each row in Table 18.1 contains the data for one benchmark. The numbers in the second and third columns are program lengths for each type of compiler. The fourth column contains the ratio of the CISC program length to the RISC program length. Averaging this ratio over all benchmarks gives the value 1.2 in the lower right. The conclusion is that CISC programs are 20% longer on average.

Chapter 18 Expectation

Benchmark	RISC	CISC	RISC/CISC
E-string search	150	120	1.25
F-bit test	120	180	0.67
Ackerman	150	300	0.5
Rec 2-sort	2800	1400	2.0
Average			1.1

**Table 18.2** The same data as in Table 18.1, but with the opposite ratio in the last column.

However, some critics of their paper took the same data and argued this way: redo the final column, taking the other ratio, RISC/CISC instead of CISC/RISC, as shown in Table 18.2.

From Table 18.2, we would conclude that RISC programs are 10% longer than CISC programs on average! We are using the same reasoning as in the paper, so this conclusion is equally justifiable—yet the result is opposite. What is going on?

**A Probabilistic Interpretation**

To resolve these contradictory conclusions, we can model the RISC vs. CISC debate with the machinery of probability theory.

Let the sample space be the set of benchmark programs. Let the random variable  $R$  be the length of the compiled RISC program, and let the random variable  $C$  be the length of the compiled CISC program. We would like to compare the average length  $\text{Ex}[R]$  of a RISC program to the average length  $\text{Ex}[C]$  of a CISC program.

To compare average program lengths, we must assign a probability to each sample point; in effect, this assigns a “weight” to each benchmark. One might like to weigh benchmarks based on how frequently similar programs arise in practice. Lacking such data, however, we will assign all benchmarks equal weight; that is, our sample space is uniform.

In terms of our probability model, the paper computes  $C/R$  for each sample point, and then averages to obtain  $\text{Ex}[C/R] = 1.2$ . This much is correct. The authors then conclude that CISC programs are 20% longer on average; that is, they conclude that  $\text{Ex}[C] = 1.2 \text{Ex}[R]$ . Therein lies the problem. The authors have implicitly used False Claim 18.5.1 to assume that  $\text{Ex}[C/R] = \text{Ex}[C]/\text{Ex}[R]$ . By using the same false logic, the critics can arrive at the opposite conclusion; namely, that RISC programs are 10% longer on average.

18.5. Expectations of Quotients

**The Proper Quotient**

We can compute  $\text{Ex}[R]$  and  $\text{Ex}[C]$  as follows:

$$\begin{aligned} \text{Ex}[R] &= \sum_{i \in \text{Range}(R)} i \cdot \Pr[R = i] \\ &= \frac{150}{4} + \frac{120}{4} + \frac{150}{4} + \frac{2800}{4} \\ &= 805, \end{aligned}$$

$$\begin{aligned} \text{Ex}[C] &= \sum_{i \in \text{Range}(C)} i \cdot \Pr[C = i] \\ &= \frac{120}{4} + \frac{180}{4} + \frac{300}{4} + \frac{1400}{4} \\ &= 500 \end{aligned}$$

Now since  $\text{Ex}[R]/\text{Ex}[C] = 1.61$ , we conclude that the *average RISC program* is 61% longer than the *average CISC program*. This is a third answer, completely different from the other two! Furthermore, this answer makes RISC look really bad in terms of code length. This one is the correct conclusion, under our assumption that the benchmarks deserve equal weight. Neither of the earlier results were correct—not surprising since both were based on the same False Claim.

**A Simpler Example**

The source of the problem is clearer in the following, simpler example. Suppose the data were as follows.

Benchmark	Processor <i>A</i>	Processor <i>B</i>	<i>B/A</i>	<i>A/B</i>
Problem 1	2	1	1/2	2
Problem 2	1	2	2	1/2
Average			1.25	1.25

Now the data for the processors *A* and *B* is exactly symmetric; the two processors are equivalent. Yet, from the third column we would conclude that Processor *B* programs are 25% longer on average, and from the fourth column we would conclude that Processor *A* programs are 25% longer on average. Both conclusions are obviously wrong.

The moral is that one must be very careful in summarizing data, we must not take an average of ratios blindly!



MIT OpenCourseWare  
<http://ocw.mit.edu>

6.042J / 18.062J Mathematics for Computer Science  
Fall 2010

For information about citing these materials or our Terms of Use, visit: <http://ocw.mit.edu/terms>.