Random Variables and Expectations

4

We have machinery to describe experiments with random outcomes, but we mostly care about numbers that are random. It is straightforward to link a number to the outcome of an experiment. The result is a random variable, a useful new idea. Random variables turn up in all sorts of places. For example, the amount of money you win or lose on a bet is a random variable. Now if you take the same bet repeatedly, you could wonder how much money will change hands in total, per bet. This yields a new and useful idea, the expected value of a random variable.

Expected values have strong properties. When one knows some expected values, you can bound various probabilities. This phenomenon parallels the property of data that we saw earlier—you don't find a large fraction of the dataset many standard deviations away from the mean. Particularly important to computer scientists (and gamblers!) is the weak law of large numbers. This law says, loosely, that the value, per bet, of repeating a bet many times will almost certainly be the expected value. Among other things, this law legitimizes estimating expectations and probabilities that are hard to calculate by using a simulation. This turns out to be really useful, because simulations are often easy programs to write and can often replace rather nasty calculations.

4.1 Random Variables

Quite commonly, we would like to deal with numbers that are random. We can do so by linking numbers to the outcome of an experiment. We define a **random variable**:

Definition 4.1 (Discrete Random Variable) Given a sample space Ω, a set of events $\mathcal{F}$, a probability function P, and a countable set of real numbers D, a discrete random variable is a function with domain Ω and range D.

This means that for any outcome ω there is a number $X(\omega)$. P will play an important role, but first we give some examples.

Example 4.1 (Numbers from Coins) We flip a coin. Whenever the coin comes up heads, we report 1; when it comes up tails, we report 0. This is a random variable.

Example 4.2 (Numbers from Coins II) We flip a coin 32 times. We record a 1 when it comes up heads, and when it comes up tails, we record a 0. This produces a 32 bit random number, which is a random variable.

D. Forsyth, *Probability and Statistics for Computer Science*,
https://doi.org/10.1007/978-3-319-64410-3_4

Example 4.3 (The Number of Pairs in a Poker Hand) We draw a hand of five cards. The number of pairs in this hand is a random variable, which takes the values 0, 1, 2 (depending on which hand we draw)

A function that takes a discrete random variable to a set of numbers is also a discrete random variable.

Example 4.4 (Parity of Coin Flips) We flip a coin 32 times. We record a 1 when it comes up heads, and when it comes up tails, we record a 0. This produces a 32 bit random number, which is a random variable. The parity of this number is also a random variable.

Associated with any value x of the random variable X are a series of events. The most important is the set of outcomes ω such that $X(\omega) = x$, which we can write $\{\omega : X(\omega) = x\}$; it is usual to simplify to $\{X = x\}$, and we will do so. The probability that a random variable X takes the value x is given by $P(\{\omega : X(\omega) = x\})$, which is more usually written $P(\{X = x\})$. This is sometimes written as $P(X = x)$, and rather often written as $P(x)$.

We could also be interested in the set of outcomes ω such that $X(\omega) \leq x$ (i.e. in $\{\omega : X(\omega) \leq x\}$), which we will write $\{X \leq x\}$; The probability that X takes a value less than or equal to x is given by $P(\{\omega : X(\omega) \leq x\})$, which is more usually written $P(\{X \leq x\})$. Similarly, we could be interested in ω such that $\{X(\omega) > x\}$, and so on.

Definition 4.2 (Probability Distribution of a Discrete Random Variable) The probability distribution of a discrete random variable is the set of numbers $P(\{X = x\})$ for each value x that X can take. The distribution takes the value 0 at all other numbers. Notice that the distribution is non-negative. The probability distribution is sometimes known as the **probability mass function**.

Definition 4.3 (Cumulative Distribution of a Discrete Random Variable) The cumulative distribution of a discrete random variable is the set of numbers $P(\{X <= x\})$ for each value x that X can take. Notice that this is a non-decreasing function of x.

Worked example 4.1 (Numbers from Coins III) We flip a biased coin 2 times. The flips are independent. The coin has $P(H) = p$, $P(T) = 1 - p$. We record a 1 when it comes up heads, and when it comes up tails, we record a 0. This produces a 2 bit random number, which is a random variable taking the values 0, 1, 2, 3. What is the probability distribution and cumulative distribution of this random variable?

Solution Probability distribution: $P(0) = (1-p)^2$; $P(1) = (1-p)p$; $P(2) = p(1-p)$; $P(3) = p^2$. Cumulative distribution: $f(0) = (1-p)^2$; $f(1) = (1-p)$; $f(2) = p(1-p) + (1-p) = (1-p^2)$; $f(3) = 1$.

Worked example 4.2 (Betting on Coins) One way to get a random variable is to think about the reward for a bet. We agree to play the following game. I flip a coin. The coin has $P(H) = p$, $P(T) = 1 - p$. If the coin comes up heads, you pay me q; if the coin comes up tails, I pay you r. The number of dollars that change hands is a random variable. What is its probability distribution?

Solution We see this problem from my perspective. If the coin comes up heads, I get q; if it comes up tails, I get $-r$. So we have $P(X = q) = p$ and $P(X = -r) = (1 - p)$, and all other probabilities are zero.

4.1.1 Joint and Conditional Probability for Random Variables

All the concepts of probability that we described for events carry over to random variables. This is as it should be, because random variables are really just a way of getting numbers out of events. However, terminology and notation change a bit.

Definition 4.4 (Joint Probability Distribution of Two Discrete Random Variables) Assume we have two random variables X and Y. The probability that X takes the value x and Y takes the value y could be written as $P(\{X = x\} \cap \{Y = y\})$. It is more usual to write it as

$$P(x, y).$$

This is referred to as the **joint probability distribution** of the two random variables (or, quite commonly, the **joint**). You can think of this as a table of probabilities, one for each possible pair of x and y values.

We will simplify notation further. Usually, we are interested in random variables, rather than potentially arbitrary outcomes or sets of outcomes. We will write $P(X)$ to denote the probability distribution of a random variable, and $P(x)$ or $P(X = x)$ to denote the probability that random variable takes a particular value. This means that, for example, the rule we could write as

$$\begin{aligned} P(\{X = x\} \,|\, \{Y = y\})P(\{Y = y\}) \\ = P(\{X = x\} \cap \{Y = y\}) \end{aligned}$$

will be written as

$$P(x|y)P(y) = P(x, y).$$

Recall the rule from Sect. 3.4.1:

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

This rule can be rewritten in our notation for random variables. This is the most familiar form of **Bayes' rule**, which is important enough to appear in its own box.

Definition 4.5 (Bayes' Rule)

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)}$$

Random variables have another useful property. If $x_0 \neq x_1$, then the event $\{X = x_0\}$ must be disjoint from the event $\{X = x_1\}$. This means that

$$\sum_x P(x) = 1$$

and that, for any y,

$$\sum_x P(x|y) = 1$$

(if you're uncertain on either of these points, check them by writing them out in the language of events).

Now assume we have the joint probability distribution of two random variables, X and Y. Recall that we write $P(\{X = x\} \cap \{Y = y\})$ as $P(x, y)$. Now consider the sets of outcomes $\{Y = y\}$ for each different value of y. These sets must be disjoint, because y cannot take two values at the same time. Furthermore, each element of the set of outcomes $\{X = x\}$ must lie in one of the sets $\{Y = y\}$. So we have

$$\sum_y P(\{X = x\} \cap \{Y = y\}) = P(\{X = x\})$$

Definition 4.6 (Marginal Probability of a Random Variable) Write $P(x, y)$ for the joint probability distribution of two random variables X and Y. Then

$$P(x) = \sum_y P(x, y) = \sum_y P(\{X = x\} \cap \{Y = y\}) = P(\{X = x\})$$

is referred to as the **marginal probability distribution** of X.

Definition 4.7 (Independent Random Variables) The random variables X and Y are **independent** if the events $\{X = x\}$ and $\{Y = y\}$ are independent for all values x and y. This means that

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

which we can rewrite as

$$P(x, y) = P(x)P(y)$$

Worked example 4.3 (Sums and Differences of Dice) You throw two dice. The number of spots on the first die is a random variable (call it X); so is the number of spots on the second die (Y). X and Y are independent. Now define $S = X + Y$ and $D = X - Y$. What is the probability distribution of S and of D?

Solution S can have values in the range $2, \ldots, 12$. There is only one way to get a $S = 2$; two ways to get $S = 3$; and so on. Using the methods of Chap. 3 for each case, the probabilities for $[2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]$ are $[1, 2, 3, 4, 5, 6, 5, 4, 3, 2, 1]/36$. Similarly, D can have values in the range $-5, \ldots, 5$. Again, using the methods of chapter Worked example 14.13, the probabilities for $[-5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5]$ are $[1, 2, 3, 4, 5, 6, 5, 4, 3, 2, 1]/36$.

Worked example 4.4 (Sums and Differences of Dice, II) Using the terminology of Example 4.3, what is the joint probability distribution of S and D?

Solution This is more interesting to display, because it's an 11×11 table. Each entry of the table represents a pair of S, D values. Many pairs can't occur (for example, for $S = 2$, D can only be zero; if S is even, then D must be even; and so on). You can work out the table by checking each case; it's in Table 4.1.

Worked example 4.5 (Sums and Differences of Dice, III) Using the terminology of Example 4.3, are X and Y independent? are S and D independent?

Solution X and Y are clearly independent. But S and D are not. There are several ways to see this. One way is to notice that, if you know $S = 2$, then you know the value of D precisely; but if you know $S = 3$, D could be either 1 or -1. This means that $P(S|D)$ depends on D, so they're not independent. Another way is to notice that the rank of the table, as a matrix, is 6, which means that it can't be the outer product of two vectors.

Table 4.1 A table of the joint probability distribution of S (vertical axis; scale $2, \ldots, 12$) and D (horizontal axis; scale $-5, \ldots, 5$) from Example 4.4

$$\frac{1}{36} \times \begin{pmatrix}
0&0&0&0&0&1&0&0&0&0&0\\
0&0&0&0&1&0&1&0&0&0&0\\
0&0&0&1&0&1&0&1&0&0&0\\
0&0&1&0&1&0&1&0&1&0&0\\
0&1&0&1&0&1&0&1&0&1&0\\
1&0&1&0&1&0&1&0&1&0&1\\
0&1&0&1&0&1&0&1&0&1&0\\
0&0&1&0&1&0&1&0&1&0&0\\
0&0&0&1&0&1&0&1&0&0&0\\
0&0&0&0&1&0&1&0&0&0&0\\
0&0&0&0&0&1&0&0&0&0&0
\end{pmatrix}$$

Worked example 4.6 (Sums and Differences of Dice, IV) Using the terminology of Example 4.3, what is $P(S|D = 0)$? what is $P(D|S = 11)$?

Solution You could work it out either of these from the table, or by first principles. If $D = 0$, S can have values $2, 4, 6, 8, 10, 12$, and each value has conditional probability $1/6$. If $S = 11$, D can have values 1, or -1, and each value has conditional probability $1/2$.

4.1.2 Just a Little Continuous Probability

Our random variables take values from a discrete set of numbers D. This makes the underlying machinery somewhat simpler to describe, and is often, but not always, enough for model building. Some phenomena are more naturally modelled as being continuous — for example, human height; human weight; the mass of a distant star; and so on. Giving a complete formal description of probability on a continuous space is surprisingly tricky, and would involve us in issues that do not arise much in practice.

These issues are caused by two interrelated facts: real numbers have infinite precision; and you can't count real numbers. A continuous random variable is still a random variable, and comes with all the stuff that a random variable comes with. We will not speculate on what the underlying sample space is, nor on the underlying events. This can all be sorted out, but requires moderately heavy lifting that isn't particularly illuminating for us. The most interesting thing for us is specifying the probability distribution. Rather than talk about the probability that a real number takes a particular value (which we can't really do satisfactorily most of the time), we will instead talk about the probability that it lies in some interval. So we can specify a probability distribution for a continuous random variable by giving a set of (very small) intervals, and for each interval providing the probability that the random variable lies in this interval.

The easiest way to do this is to supply a **probability density function**. Let $p(x)$ be a probability density function (often called a **pdf** or **density**) for a continuous random variable X. We interpret this function by thinking in terms of small intervals. Assume that dx is an infinitesimally small interval. Then

$$\begin{aligned} p(x)dx = P(\{&\text{event that } X \text{ takes a value in} \\ &\text{the range } [x, x + dx]\}). \end{aligned}$$

Important properties of probability density functions follow from this definition.

Useful Facts 4.1 (Properties of Probability Density Functions)

- Probability density functions are non-negative. This follows from the definition; a negative value at some u would imply that $P(\{x \in [u, u+du]\})$ was negative, and this cannot occur.
- For $a < b$

$$P(\{X \text{ takes a value in the range } [a, b]\}) = \int_a^b p(x)dx.$$

which we obtain by summing $p(x)dx$ over all the infinitesimal intervals between a and b.
- We must have that

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

This is because

$$P(\{X \text{ takes a value in the range } [-\infty, \infty]\}) = 1 = \int_{-\infty}^{\infty} p(x)dx$$

The property that

$$\int_{-\infty}^{\infty} p(x)dx = 1$$

is useful, because when we are trying to determine a probability density function, we can ignore a constant factor. So if $g(x)$ is a non-negative function that is proportional to the probability density function (often pdf) we are interested in, we can recover the pdf by computing

$$p(x) = \frac{1}{\int_{-\infty}^{\infty} g(x)dx} g(x).$$

This procedure is sometimes known as **normalizing**, and $\int_{-\infty}^{\infty} g(x)dx$ is the **normalizing constant**.

One good way to think about pdf's is as the limit of a histogram. Imagine you collect an arbitrarily large dataset of data items, each of which is independent. You build a histogram of that dataset, using arbitrarily narrow boxes. You scale the histogram so that the sum of the box areas is one. The result is a probability density function.

The pdf doesn't represent the probability that a random variable takes a value. Instead, you should think of $p(x)$ as being the limit of a ratio (which is why it's called a density):

$$\frac{\text{the probability that the random variable will lie in a small interval centered on } x}{\text{the length of the small interval centered on } x}$$

Notice that, while a pdf has to be non-negative, and it has to integrate to 1, it does *not* have to be smaller than one. A ratio like this could be a lot larger than one, as long as it isn't larger than one for too many x (because the integral must be one). In fact, probability density functions can be strange functions (exercises).

Worked example 4.7 (A Probability Density Function that is Larger than One) Assume we have a physical system that can produce random numbers. It produces numbers in the range 0 to ϵ, where $\epsilon > 0$. Each number has the same probability of appearing. No number larger than ϵ or smaller than 0 can ever appear. What is the probability density function?

(continued)

Solution Write $p(x)$ for the probability density function. We must have that $p(x) = 0$ for $x < 0$ and $p(x) = 0$ for $x > \epsilon$. We must have that $p(x)$ is constant between 0 and ϵ and that

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

So

$$p(x) = \begin{cases} 0 & \text{if } x < 0 \\ 0 & \text{if } x > \epsilon \\ \frac{1}{\epsilon} & \text{otherwise} \end{cases}$$

Notice that if $\epsilon < 1$, we have that $p(x) > 1$ for all x.

Remember this: *Probability notation can be quirky. Usually, one uses a big P for actual probabilities, and a small p for probability densities. The argument, or context, is supposed to tell you which probability distribution is meant (i.e $P(X)$ likely refers to a different probability distribution than $P(Y)$, which should strike a computer scientist familiar with dummy variables as bizarre). Because the probability distribution for a discrete random variable is a collection of probabilities, following this convention requires that such a probability distribution be written with a big P. However, having different notation for discrete and continuous random variables can get quite clunky. In application areas it is usual to write a small p for a probability distribution, and whether a density or a distribution is intended depends on whether the random variable is continuous or discrete. However, if you want to emphasize that a probability is intended, you can write P. I will follow this convention. To add to the fun, you may encounter $p(x)$ with the meaning "some probability distribution" or $p(x)$ meaning "the value of the probability distribution $P(\{X = x\})$ at the point x" or $p(x)$ with the meaning "the probability distribution $P(\{X = x\})$ as a function of x". You can usually figure out what is intended as long as you don't think too closely about it (authors are often quite inconsistent); context may help disambiguate different intended meanings, too. Cumulative distributions are often written with an f, so that an unexpected $f(x)$ might mean $P(\{X <= x\})$.*

4.2 Expectations and Expected Values

Example 4.2 described a simple game. I flip a coin. The coin has $P(H) = p$, $P(T) = 1 - p$. If the coin comes up heads, you pay me q; if the coin comes up tails, I pay you r. Now imagine we play this game many times. Our frequency definition of probability means that in N games, we expect to see about pN heads and $(1 - p)N$ tails. In turn, this means that my total income from these N games should be about $(pN)q - ((1 - p)N)r$. The N in this expression is inconvenient; instead, we could say that for any single game, my expected income is

$$pq - (1 - p)r.$$

This isn't the actual income from a single game (which would be either q or $-r$, depending on what the coin did). Instead, it's an estimate of what would happen over a large number of games, on a per-game basis. This is an example of an expected value.

4.2.1 Expected Values

Definition 4.8 (Expected Value) Given a discrete random variable X which takes values in the set $\mathcal{D}$ and which has probability distribution P, we define the expected value

(continued)

$$\mathbb{E}[X] = \sum_{x \in \mathcal{D}} xP(X = x).$$

This is sometimes written $\mathbb{E}_P[X]$, to clarify which distribution one has in mind.

Notice that an expected value could take a value that the random variable doesn't take.

Example 4.5 (Betting on Coins) We agree to play the following game. I flip a fair coin (i.e. $P(H) = P(T) = 1/2$). If the coin comes up heads, you pay me 1; if the coin comes up tails, I pay you 1. The expected value of my income is 0, even though the random variable never takes that value.

Worked example 4.8 (Betting on Coins, Again) We agree to play the following game. I flip a fair coin (i.e. $P(H) = P(T) = 1/2$). If the coin comes up heads, you pay me 2; if the coin comes up tails, I pay you 1. What is the expected value of this game?

Solution The expected value of my income is

$$(\frac{1}{2}) \times 2 - (\frac{1}{2}) \times 1 = \frac{1}{2}.$$

Notice this isn't even an integer, and there's no way that any one instance of the game would yield a payoff of $1/2$. But this is what I would get, per game, if I played many times.

Your intuition is likely to tell you that the game of Example 4.8 is good for me and bad for you. This intuition is correct. It turns out that an even stronger statement is possible: playing this game repeatedly is pretty much guaranteed to be excellent for me and disastrous for you. It'll take some pages before I can be crisp about precisely what I mean here and why it is true.

Definition 4.9 (Expectation) Assume we have a function f that maps a discrete random variable X into a set of numbers $\mathcal{D}_f$. Then $f(X)$ is a discrete random variable, too, which we write F. The expected value of this random variable is written

$$\mathbb{E}[f] = \sum_{u \in \mathcal{D}_f} uP(F = u) = \sum_{x \in \mathcal{D}} f(x)P(X = x)$$

which is sometimes referred to as "the expectation of f". The process of computing an expected value is sometimes referred to as "taking expectations". This is sometimes written $\mathbb{E}_P[f]$ or even $\mathbb{E}_{P(X)}[f]$, to clarify which distribution one has in mind.

We can compute expectations for continuous random variables, too, though summing over all values now turns into an integral. Assume I have a continuous random variable X with probability density function $p(x)$. Remember I interpret the probability density function as meaning that, for an infinitesimal interval size dx, $p(x)dx = P(\{X \in [x, x + dx]\})$). Divide the set of possible values that X can take into small intervals of width Δx, centered on x_i. We can construct a discrete random variable $\hat{X}$ which takes values x_i. We have that $P(\{\hat{X} = x_i\}) \approx p(x_i)\Delta x$, where I used the approximation sign because Δx may not be infinitesimally small.

Now write $\mathbb{E}\left[\hat{X}\right]$ for the expected value of $\hat{X}$. We have

$$\mathbb{E}\left[\hat{X}\right] = \sum_{x_i} x_i P(x_i) \approx \sum_{x_i} x_i p(x_i) \Delta x.$$

As the intervals limit to infinitesimal intervals, $\hat{X}$ limits to X (think of a picture of a histogram with infinitely narrow boxes). Then $\mathbb{E}\left[\hat{X}\right]$ has a limit which is an integral, and this defines the expected value. So we have the expressions in the boxes below.

Definition 4.10 (Expected Value of a Continuous Random Variable) Given a continuous random variable X which takes values in the set $\mathcal{D}$ and which has probability distribution P, we define the expected value

$$\mathbb{E}[X] = \int_{x \in \mathcal{D}} xp(x)dx.$$

This is sometimes written $\mathbb{E}_p[X]$, to clarify which distribution one has in mind.

The expected value of a continuous random variable could be a value that the random variable doesn't take, too. Notice one attractive feature of the $\mathbb{E}[X]$ notation; we don't need to make any commitment to whether X is a discrete random variable (where we would write a sum) or a continuous random variable (where we would write an integral). The reasoning by which we turned a sum into an integral works for functions of continuous random variables, too.

Definition 4.11 (Expectation of a Continuous Random Variable) Assume we have a function f that maps a continuous random variable X into a set of numbers $\mathcal{D}_f$. Then $f(X)$ is a continuous random variable, too, which we write F. The expected value of this random variable is

$$\mathbb{E}[f] = \int_{x \in \mathcal{D}} f(x)p(x)dx$$

which is sometimes referred to as "the expectation of f". The process of computing an expected value is sometimes referred to as "taking expectations".

Under some circumstances the expected value may not exist. The integral needs to exist, and be finite, for us to interpret the expected value meaningfully, and that isn't guaranteed for every continuous random variable. Nothing we do will encounter this issue, and so we will ignore it.

You can see an expectation as an operation you apply to a random variable. It doesn't matter whether the random variable is discrete or continuous; that just changes the recipe for computing the value of the expectation. The crucial property of this operation is that it is linear; this is so important I have put it in its own box.

Useful Facts 4.2 (Expectations Are Linear)
Write f, g for functions of random variables.

- $\mathbb{E}[0] = 0$
- for any constant k, $\mathbb{E}[kf] = k\mathbb{E}[f]$
- $\mathbb{E}[f + g] = \mathbb{E}[f] + \mathbb{E}[g]$.

I have written this box in a rather compact form. This is because the expression $\mathbb{E}[X]$ for the expected value of a random variable is actually a special case of $\mathbb{E}[f]$—one just uses the identity function for f. So the box also tells us that $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$, and so on.

4.2.2 Mean, Variance and Covariance

There are three very important expectations with special names.

Definition 4.12 (Mean or Expected Value) The mean or expected value of a random variable X is

$$\mathbb{E}[X]$$

Worked example 4.9 (Mean of a Coin Flip) We flip a biased coin, with $P(H) = p$. The random variable X has value 1 if the coin comes up heads, 0 otherwise. What is the mean of X? (i.e. $\mathbb{E}[X]$).

Solution $\mathbb{E}[X] = \sum_{x \in D} xP(X = x) = 1p + 0(1 - p) = p$

Definition 4.13 (Variance) The variance of a random variable X is

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

Useful Facts 4.3 (Properties of Variance)
We have:

- For any constant k, $\mathsf{var}[k] = 0$;
- $\mathsf{var}[X] \geq 0$;
- $\mathsf{var}[kX] = k^2\mathsf{var}[X]$;
- and, if X and Y are independent, then $\mathsf{var}[X + Y] = \mathsf{var}[X] + \mathsf{var}[Y]$.

The first three are obvious, and the fourth appears in the exercises.

Useful Facts 4.4 (Variance, a Useful Expression)

$$\begin{aligned}\mathsf{var}[X] &= \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \\ &= \mathbb{E}\left[(X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2)\right] \\ &= \mathbb{E}\left[X^2\right] - 2\mathbb{E}[X]\mathbb{E}[X] + \mathbb{E}[X]^2 \\ &= \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2\end{aligned}$$

Worked example 4.10 (Variance of a Coin Flip) We flip a biased coin, with $P(H) = p$. The random variable X has value 1 if the coin comes up heads, 0 otherwise. What is the variance of X? (i.e. $\mathsf{var}[X]$).

Solution $\mathsf{var}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = (1p - 0(1 - p)) - p^2 = p(1 - p)$

Worked example 4.11 (Variance) Can a random variable have $\mathbb{E}[X] > \sqrt{\mathbb{E}[X^2]}$?

Solution No, because that would mean that $\mathbb{E}\left[(X - \mathbb{E}[X])^2\right] < 0$. But this is the expected value of a non-negative quantity; it must be non-negative.

Worked example 4.12 (More Variance) We just saw that a random variable can't have $\mathbb{E}[X] > \sqrt{\mathbb{E}[X^2]}$. But I can easily have a random variable with large mean and small variance—isn't this a contradiction?

Solution No, you're confused. Your question means you think that the variance of X is given by $\mathbb{E}\left[X^2\right]$; but actually $\mathsf{var}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$

Now assume that we have a probability distribution $P(X)$ defined on some discrete set of numbers. There is some random variable that produced this probability distribution. This means that we could talk about the mean of a probability distribution P (rather than the mean of a random variable whose probability distribution is $P(X)$). It is quite usual to talk about the mean of a probability distribution. Furthermore, we could talk about the variance of a probability distribution P (rather than the variance of a random variable whose probability distribution is $P(X)$).

Definition 4.14 (Covariance) The covariance of two random variables X and Y is

$$\mathsf{cov}\,(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

Useful Facts 4.5 (Covariance, Useful Expression)

$$\begin{aligned}\mathsf{cov}\,(X, Y) &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\ &= \mathbb{E}[(XY - Y\mathbb{E}[X] - X\mathbb{E}[Y] \\ &\qquad + \mathbb{E}[X]\mathbb{E}[Y])] \\ &= \mathbb{E}[XY] - 2\mathbb{E}[Y]\mathbb{E}[X] + \mathbb{E}[X]\mathbb{E}[Y] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].\end{aligned}$$

Useful Facts 4.6 (Independent Random Variables Have Zero Covariance)
We have:

- if X and Y are independent, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$;
- if X and Y are independent, then $\mathsf{cov}\,(X, Y) = 0$.

If the first is true, then the second is obviously true (apply the expression of useful facts 4.5).

Proposition *If X and Y are independent random variables, then* $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.

Proof Recall that $\mathbb{E}[X] = \sum_{x \in D} xP(X = x)$, so that

$$\begin{aligned}
\mathbb{E}[XY] &= \sum_{(x,y)\in D_x \times D_y} xyP(X = x, Y = y) \\
&= \sum_{x\in D_x} \sum_{y \in D_y} (xyP(X = x, Y = y)) \\
&= \sum_{x\in D_x} \sum_{y \in D_y} (xyP(X = x)P(Y = y)) \\
&\quad \text{because } X \text{ and } Y \text{ are independent} \\
&= \sum_{x\in D_x} \sum_{y \in D_y} (xP(X = x))\,(yP(Y = y)) \\
&= \left(\sum_{x\in D_x} xP(X{=}x)\right)\left(\sum_{y\in D_y} yP(Y{=}y)\right) \\
&= (\mathbb{E}[X])(\mathbb{E}[Y]).
\end{aligned}$$

This is certainly not true when X and Y are not independent (try $Y = -X$).

Useful Facts 4.7 (Variance as Covariance)
We have

$$\mathsf{var}[X] = \mathsf{cov}\,(X, X)$$

(substitute into definitions).

The variance of a random variable is often inconvenient, because its units are the square of the units of the random variable. Instead, we could use the **standard deviation**.

Definition 4.15 (Standard Deviation) The **standard deviation** of a random variable X is defined as

$$\mathsf{std}\,(\{X\}) = \sqrt{\mathsf{var}[X]}$$

You do need to be careful with standard deviations. If X and Y are independent random variables, then $\mathsf{var}[X + Y] = \mathsf{var}[X] + \mathsf{var}[Y]$, but $\mathsf{std}\,(\{X + Y\}) = \sqrt{\mathsf{std}\,(\{X\})^2 + \mathsf{std}\,(\{Y\})^2}$. One way to avoid getting mixed up is to remember that variances add, and derive expressions for standard deviations from that.

4.2.3 Expectations and Statistics

I have now used each of the terms mean, variance, covariance, and standard deviation in two slightly different ways. One sense of each term, expounded in Sect. 1.3, describes a property of a dataset. These are known as **descriptive statistics**. The other sense, described above, is a property of probability distributions. These are known as **expectations**. The reason we use one name for two notions is that the notions are not really all that different.

Here is a useful construction to illustrate the point. Imagine we have a dataset $\{x\}$ of N items, where the i'th item is x_i. Build a random variable X using this dataset by placing the same probability on each data item. This means that each data item has probability $1/N$. Write $\mathbb{E}[X]$ for the mean of this distribution. We have

$$\mathbb{E}[X] = \sum_i x_i P(x_i) = \frac{1}{N}\sum_i x_i = \mathsf{mean}\left(\{x\}\right)$$

and, by the same reasoning,

$$\mathsf{var}[X] = \mathsf{var}\left(\{x\}\right).$$

This construction works for standard deviation and covariance, too. For this particular distribution (sometimes called the **empirical distribution**), the expectations have the same value as the descriptive statistics.

In Sect. 4.3.4, we will see a form of converse to this fact. Imagine we have a dataset that consists of independent, identically distributed samples from a probability distribution (i.e. we know that each data item was obtained independently from the distribution). For example, we might have a count of heads in each of a number of coin flip experiments. Then the descriptive statistics will turn out to be accurate estimates of the expectations.

4.3 The Weak Law of Large Numbers

Assume you see repeated values of a random variable. For example, let X be the random variable which has value 1 if a coin comes up heads (which happens with probability p) and -1 if it comes up tails. You now actually flip a coin N times, recording 1 for heads and -1 for tails. Intuition should say that the average of these numbers should be a good estimate of the value of $\mathbb{E}[X]$, by the following argument. You should see 1 about pN times, and -1 about $(1-p)N$ times. So the average should be close to $p-(1-p)$, which is $\mathbb{E}[X]$. Furthermore, intuition should suggest that this estimate gets better as the number of flips goes up.

These intuitions are correct. You can estimate expectations accurately by experiment. This is extremely useful, because it means that you can use quite simple programs to estimate values that might take a great deal of work to obtain any other way. Most people find it natural that something of this sort should be true. What is neat is that it is quite easy to prove.

4.3.1 IID Samples

We need first to be crisp about what we are averaging. Imagine a random variable X, obtained by flipping a fair coin and reporting 1 for an H and -1 for a T. We can talk about the probability distribution $P(X)$ of this random variable; we can talk about the expected value that the random variable takes; but the random variable itself doesn't have a value. However, if we actually flip a coin, we get either a 1 or a -1. Observing a value is sometimes called a **trial**. The resulting value is often called a **sample** of the random variable (or of its probability distribution); it is sometimes called a **realization**. So flipping the coin is a trial, and the number you get is a sample. If we flipped a coin many times, we'd have a set of numbers (or samples). These numbers would be independent. Their histogram would look like $P(X)$. Collections of data items like this are important enough to have their own name.

Assume we have a set of data items x_i such that (a) they are independent; and (b) the histogram of a very large set of data items looks increasingly like the probability distribution $P(X)$ as the number of data items increases. Then we refer to these data items as **independent identically distributed samples** of $P(X)$; for short, **iid samples** or even just **samples**. It's worth knowing that it can be a difficult computational problem to get IID samples from some given probability distribution. For all of the cases we will deal with, it will be obvious how to get IID samples. Usually, they were generated for us—i.e. somebody flipped the coin, etc.

Now assume you take N IID samples, and average them. The weak law of large numbers states that, as N gets larger, this average is an increasingly good estimate of $\mathbb{E}[X]$. This fact allows us estimate expectations (and so probabilities) by simulation. Furthermore, it will allow us to make strong statements about how repeated games with random outcomes will behave. Finally, it will allow us to build a theory of decision making.

4.3.2 Two Inequalities

To go further, we need two useful inequalities. Consider

$$\mathbb{E}[|X|] = \sum_{x \in D} |x| P(\{X = x\}).$$

Now notice that all the terms in the sum are non-negative. Then the only way to have a small value of $\mathbb{E}[|X|]$ is to be sure that, when $|x|$ is large, $P(\{X = x\})$ is small. It turns out to be possible (and useful!) to be more crisp about how quickly $P(\{X = x\})$ falls as $|x|$ grows, resulting in Markov's inequality (which I'll prove below)

Definition 4.16 (Markov's Inequality) **Markov's inequality** is

$$P(\{|X| \geq a\}) \leq \frac{\mathbb{E}[|X|]}{a}.$$

Notice that we've seen something like this before (the result about standard deviation in Sect. 1.3.2 has this form). The reason this is worth proving is that it leads to a second result, and that gives us the weak law of large numbers. It should seem clear that the probability of a random variable taking a particular value must fall off rather fast as that value moves away from the mean, in units scaled to the standard deviation. This is because values of a random variable that are many standard deviations above the mean must have low probability, otherwise the values would occur more often and so the standard deviation would be bigger. This result is Chebyshev's inequality, which I shall also prove below.

Definition 4.17 (Chebyshev's Inequality) **Chebyshev's inequality** is

$$P(\{|X - \mathbb{E}[X]| \geq a\}) \leq \frac{\mathsf{var}[X]}{a^2}.$$

It is common to see this in another form, obtained by writing σ for the standard deviation of X, substituting $k\sigma$ for a, and rearranging

$$P(\{|X - \mathbb{E}[X]| \geq k\sigma\}) \leq \frac{1}{k^2}$$

We care about Chebyshev's inequality because it gives us the weak law of large numbers.

4.3.3 Proving the Inequalities

An **indicator function** is a function that is one when some condition is true, and zero otherwise. The reason indicator functions are useful is that their expected values have interesting properties.

Definition 4.18 (Indicator Functions) An indicator function for an event is a function that takes the value zero for values of x where the event does not occur, and one where the event occurs. For the event $\mathcal{E}$, we write

$$\mathbb{I}_{[\mathcal{E}]}(x)$$

for the relevant indicator function.

I used a small x in the definition, because this is a function; the argument doesn't *need* to be a random variable. You should think about an indicator function as testing the value of its argument to tell whether it lies in the event or not, and reporting 1 or 0 accordingly. For example,

$$\mathbb{I}_{[\{|x|\}\leq a]}(x) = \begin{cases} 1 \text{ if } -a < x < a \\ 0 \quad \text{otherwise} \end{cases}$$

Indicator functions have one useful property.

$$\mathbb{E}_P\left[\mathbb{I}_{[\mathcal{E}]}\right] = P(\mathcal{E})$$

which you can establish by checking the definition of expectations.

Proposition *Markov's inequality: for X a random variable, $a > 0$,*

$$P(\{|X| \geq a\}) \leq \frac{\mathbb{E}[|X|]}{a}.$$

Proof (from Wikipedia). Notice that, for $a > 0$,

$$a\mathbb{I}_{[\{|X|\geq a\}]}(X) \leq |X|$$

(because if $|X| \geq a$, the LHS is a; otherwise it is zero). Now we have

$$\mathbb{E}\left[a\mathbb{I}_{[\{|X|\geq a\}]}\right] \leq \mathbb{E}[|X|]$$

but, because expectations are linear, we have

$$\mathbb{E}\left[a\mathbb{I}_{[\{|X|\geq a\}]}\right]=a\mathbb{E}\left[\mathbb{I}_{[\{|X|\geq a\}]}\right]=aP(\{|X|\geq a\})$$

and so we have

$$aP(\{|X| \geq a\}) \leq \mathbb{E}[|X|]$$

and we get the inequality by division, which we can do because $a > 0$.

Proposition *Chebyshev's inequality: for X a random variable, $a > 0$,*

$$P(\{|X - \mathbb{E}[X]| \geq a\}) \leq \frac{\mathsf{var}[X]}{a^2}.$$

Proof Write U for the random variable $(X - \mathbb{E}[X])^2$. Markov's inequality gives us

$$P(\{|U| \geq w\}) \leq \frac{\mathbb{E}[|U|]}{w}$$

Now notice that, if $w = a^2$,

$$P(\{|U| \geq w\}) = P(\{|X - \mathbb{E}[X]| \geq a\})$$

so we have

$$\begin{aligned} P(\{|U| \geq w\}) &= P(\{|X - \mathbb{E}[X]| \geq a\}) \\ &\leq \frac{\mathbb{E}[|U|]}{w} = \frac{\mathsf{var}[X]}{a^2} \end{aligned}$$

4.3.4 The Weak Law of Large Numbers

Assume we have a set of N IID samples x_i of a probability distribution $P(X)$. Write

$$X_N = \frac{\sum_{i=1}^{N} x_i}{N}.$$

Now X_N is a random variable (the x_i are IID samples, and for a different set of samples you will get a different, random, X_N). Notice that $P(X = x_1, X = x_2, \ldots, X = x_n) = P(X = x_1)P(X = x_2) \ldots P(X = x_n)$, because the samples are independent and each is a sample of $P(X)$. This means that

$$\mathbb{E}[X_N] = \mathbb{E}[X]$$

because

$$\mathbb{E}[X_N] = \left(\frac{1}{N}\right)\sum_{i=1}^{N}\mathbb{E}[X].$$

This means that

$$\frac{\sum_{i=1}^{N} x_i}{N}$$

should be an accurate estimate of $\mathbb{E}[X]$. The weak law of large numbers states that, as N gets large, the estimate becomes more accurate.

Definition 4.19 (Weak Law of Large Numbers) If $P(X)$ has finite variance, then for any positive number ϵ

$$\lim_{N\to\infty} P(\{|X_N - \mathbb{E}[X]| \geq \epsilon\}) = 0.$$

Equivalently, we have

$$\lim_{N\to\infty} P(\{|X_N - \mathbb{E}[X]| < \epsilon\}) = 1.$$

Proposition *Weak law of large numbers*

$$\lim_{N\to\infty} P(\{|X_N - \mathbb{E}[X]| \geq \epsilon\}) = 0.$$

Proof Write $\mathsf{var}(\{X\}) = \sigma^2$. Choose $\epsilon > 0$. Now we have that

$$\begin{aligned}\mathsf{var}(\{X_N\}) &= \mathsf{var}\left(\left\{\frac{\sum_{i=1}^{N} x_i}{N}\right\}\right)\\ &= (\frac{1}{N^2})\mathsf{var}\left(\left\{\sum_{i=1}^{N} x_i\right\}\right)\\ &= (\frac{1}{N^2})(N\sigma^2)\\ &\quad\text{the } x_i \text{ are independent}\\ &= \frac{\sigma^2}{N}\end{aligned}$$

and that

$$\mathbb{E}[X_N] = \mathbb{E}[X].$$

(continued)

Now Chebyshev's inequality gives

$$P(\{|X_N - \mathbb{E}[X]| \geq \epsilon\}) \leq \frac{\sigma^2}{N\epsilon^2}$$

so

$$\lim_{N\to\infty} P(\{|X_N - \mathbb{E}[X]| \geq \epsilon\}) = \lim_{N\to\infty} \frac{\sigma^2}{N\epsilon^2} = 0.$$

The weak law of large numbers gives us a very valuable way of thinking about expectations. Assume we have a random variable X. Then the weak law says that, if you observe a large number of IID samples of this random variable, the average of the values you observe should be very close to $\mathbb{E}[X]$. This result is extremely powerful. The next section explores some applications. The weak law allows us to estimate expectations (and so probabilities, which are expectations of indicator functions) by observing random behavior. The weak law can be used to build a theory of decision making.

4.4 Using the Weak Law of Large Numbers

4.4.1 Should You Accept a Bet?

We can't answer this as a moral question, but we can as a practical question, using expectations. Generally, a bet involves an agreement that amounts of money will change hands, depending on the outcome of an experiment. Mostly, you are interested in how much you get from the bet, so it is natural to give sums of money you receive a positive sign, and sums of money you pay out a negative sign. The weak law says that if you repeat a bet many times, you are increasingly likely to receive the expected value of the bet, per bet. Under this convention, the practical answer is easy: accept a bet enthusiastically if its expected value is positive, otherwise decline it. It is interesting to notice how poorly this advice describes actual human behavior.

Worked example 4.13 (Red or Black?) On a roulette wheel (see p. xxiii if you can't remember how these work), you can bet on (among other things) whether a red number or a black number comes up. If you bet 1 on red, and a red number comes up, you keep your stake and get 1; if a black number or a zero comes up, you get -1 (i.e. the house keeps your bet). What is the expected value of a bet of 1 on a wheel with one, two and three zeros?

Solution Write p_r for the probability a red number comes up. The expected value is $1 \times p_r + (-1)(1 - p_r)$ which is $2p_r - 1$. For one zero, $p_r =$ (number of red numbers)/ (total number of numbers) $= 18/37$. So the expected value is $-1/37$ (you lose about three cents each time you bet a dollar). For two zeros, $p_r = 18/38$. So the expected value is $-2/38 = -1/19$ (you lose slightly more than five cents each time you bet a dollar). For three zeros, $p_r = 18/39$. So the expected value is $-3/39 = -1/13$ (you lose slightly less than eight cents each time you bet a dollar).

Notice that in the roulette game, the money you lose will go to the house. So the expected value to the house is just the negative of the expected value to you. You might not play the wheel often, but the house plays the wheel very often when there are many players. The weak law means a house with many players can rely on receiving about three, five, or eight cents per dollar bet, depending on the number of zeros on the wheel. This is a partial explanation of why there are lots of roulette wheels, and usually free food nearby. Not all bets are like this, though.

Worked example 4.14 (Coin Game) In this game, P1 flips a fair coin and P2 calls "H" or "T". If P2 calls right, then P1 throws the coin into the river; otherwise, P1 keeps the coin. The coin belongs to P1, and has value 1. What is the expected value of this game to P2? and to P1?

(continued)

Solution To P2, which we do first, because it's easiest: P2 gets 0 if P2 calls right, and 0 if P2 calls wrong; these are the only cases, so the expected value is 0. To P1: P1 gets -1 if P2 calls right, and 0 if P1 calls wrong. The coin is fair, so the probability P2 calls right is 1/2. The expected value is $-1/2$. While I can't explain why people would play such a game, I've actually seen this done.

We call a bet **fair** when its expected value is zero. Taking a bet with a negative expected value is unwise, because, on average, you will lose money. Worse, the more times you play, the more you lose. Similarly, repeatedly taking a bet with a positive expected value is reliably profitable. However, you do need to be careful you computed the expected value right.

Worked example 4.15 (Birthdays in Succession) P1 and P2 agree to the following bet. P1 gives P2 a stake of 1. If three people, stopped at random on the street, have birthdays in succession (i.e. Mon-Tue-Wed, and so on), then P2 gives P1 100. Otherwise, P1 loses the stake. What is the expected value of this bet to P1?

Solution Write p for the probability of winning. Then the expected value is $p \times 100 - (1-p) \times 1$. We computed p in Example 3.45 (it was $1/49$). So the bet is worth $(52/49)$, or slightly more than a dollar, to P1. P1 should be happy to agree to this as often as possible.

The reason P2 agrees to bets like that of Example 4.15 is most likely that P2 can't compute the probability exactly. P2 thinks the event is quite unlikely, so the expected value is negative; but it isn't as unlikely as P2 thought it was, and this is how P1 makes a profit. This is one of the many reasons you should be careful accepting a bet from a stranger: they might be able to compute better than you.

4.4.2 Odds, Expectations and Bookmaking: A Cultural Diversion

Gamblers sometimes use a terminology that is a bit different from ours. In particular, the term **odds** is important. The term comes from the following idea: P1 pays a bookmaker b (the stake) to make a bet; if the bet is successful, P1 receives a and the stake back, and if not, loses the original stake. This bet is referred to as odds of $a : b$ (read "odds of a to b").

Assume the bet is fair, so that the expected value is zero. Write p for the probability of winning. The net income to P1 is $ap-b(1-p)$. If this is zero, then $p = b/(a+b)$. So you can interpret odds in terms of probability, *if* you assume the bet is fair.

A bookmaker sets odds at which to accept bets from gamblers. The bookmaker does not wish to lose money at this business, and so must set odds which are potentially profitable. Doing so is not simple (bookmakers can, and occasionally do, lose catastrophically, and go out of business). In the simplest case, assume that the bookmaker knows the probability p that a particular bet will win. Then the bookmaker could set odds of $(1-p)/p : 1$. In this case, the expected value of the bet is zero; this is fair, but not attractive business, so the bookmaker will set odds assuming that the probability is a bit higher than it really is. There are other bookmakers out there, so there is some reason for the bookmaker to try to set odds that are close to fair.

In some cases, you can tell when you are dealing with a bookmaker who is likely to go out of business soon. For example, imagine there are two horses running in a race, both at 10 : 1 odds—whatever happens, you could win by betting 1 on each. There is a more general version of this phenomenon. Assume the bet is placed on a horse race, and that bets pay off only for the winning horse. Assume also that exactly one horse will win (i.e. the race is never scratched, there aren't any ties, etc.), and write the probability that the i'th horse will win as p_i. Then

$$\sum_{i \in \text{horses}} p_i$$

must be 1. Now if the bookmaker's odds yield a set of probabilities that is less than 1, their business should fail, because there is at least one horse on which they are paying out too much. Bookmakers deal with this possibility by writing odds so that $\sum_{i \in \text{horses}} p_i$ is larger than one.

But this is not the only problem a bookmaker must deal with. The bookmaker doesn't actually know the probability that a particular horse will win, and must account for errors in this estimate. One way to do so is to collect as much information as possible (talk to grooms, jockeys, etc.). Another is to look at the pattern of bets that have been placed already. If the bookmaker and the gamblers agree on the probability that each horse will win, then there should be no expected advantage to choosing one horse over another—each should pay out slightly less than zero to the gambler (otherwise the bookmaker doesn't eat). But if the bookmaker has underestimated the probability that a particular horse will win, a gambler may get a positive expected payout by betting on that horse. This means that if one particular horse attracts a lot of money from bettors, it is wise for the bookmaker to offer less generous odds on that horse. There are two reasons: first, the bettors might know something the bookmaker doesn't, and they're signalling it; second, if the bets on this horse are very large and it wins, the bookmaker may not have enough capital left to pay out or to stay in business. All this means that real bookmaking is a complex, skilled business.

4.4.3 Ending a Game Early

Imagine two people are playing a game for a stake, but must stop early—who should get what percentage of the stake? One way to do this is to give each player what they put in at the start, but this is (mildly) unfair if one has an advantage over the other. The alternative is to give each player the expected value of the game at that state for that player. Sometimes one can compute that expectation quite easily.

Worked example 4.16 (Ending a Game Early) Two players each pay 25 to play the following game. They toss a fair coin. If it comes up heads, player H wins that toss; if tails, player T wins. The first player to reach 10 wins takes the stake of 50. But one player is called away when the state is 8–7 (H-T)—how should the stake be divided?

Solution In this state, each player can either win—and so get 50—or lose—and so get 0. The expectation for H is $50P(\{\text{H wins from 8-7}\}) + 0P(\{\text{T wins from 8-7}\})$, so we need to compute $P(\{\text{H wins from 8-7}\})$. Similarly, the expectation for T is $50P(\{\text{T wins from 8-7}\}) + 0P(\{\text{H wins from 8-7}\})$, so we need to compute $P(\{\text{T wins from 8-7}\})$; but $P(\{\text{T wins from 8-7}\}) = 1 - P(\{\text{H wins from 8-7}\})$. Now it is slightly easier to compute $P(\{\text{T wins from 8-7}\})$, because T can only win in two ways: 8–10 or 9–10. These are independent. For T to win 8–10, the next three flips must come up T, so that event has probability $1/8$. For T to win 9–10, the next four flips must have one H in them, but the last flip may not be H (or else H wins); so the next four flips could be HTTT, THTT, or TTHT. The probability of this is $3/16$. This means the total probability that T wins is $5/16$. So T should get 15.625 and H should get the rest (although they might have to flip for the odd half cent).

4.4.4 Making a Decision with Decision Trees and Expectations

Imagine we have to choose an action. Once we have chosen, a sequence of random events occurs, and we get a reward with some probability. Which action should we choose? A good answer is to choose the action with the best expected outcome. If we encounter this situation repeatedly, the weak law tells us that choosing any other action than the one with best expected outcome is unwise. If we make a choice that is even only slightly worse than the best, we will reliably do worse than we could. This is a very common recipe, and it can be applied to many situations. Usually, but not always, the reward is in money, and we will compute with money rewards for the first few examples.

For such problems, it can be useful to draw a **decision tree**. A decision tree is a drawing of possible outcomes of decisions, which makes costs, benefits and random elements explicit. Each node of the tree represents a test of an attribute (which could be either a decision, or a random variable), and each edge represents a possible outcome of a test. The final outcomes are leaves. Usually, decision nodes are drawn as squares, chance elements as circles, and leaves as triangles.

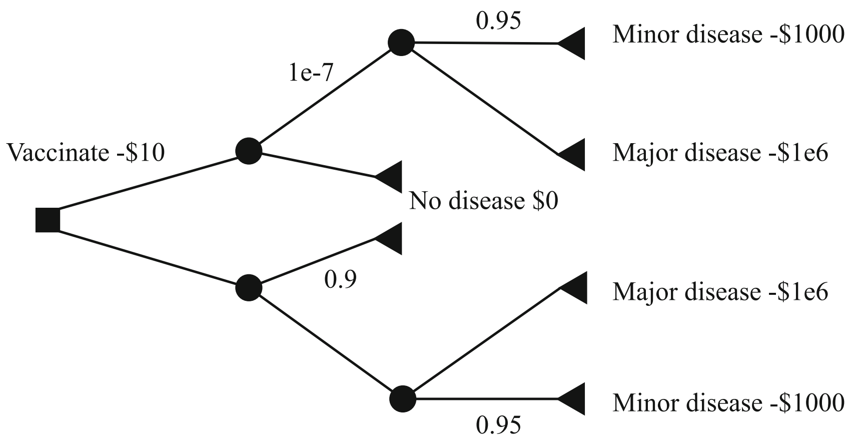

Fig. 4.1 A decision tree for the vaccination problem. The only decision is whether to vaccinate or not (the box at the root of the tree). I have only labelled edges where this is essential, so I did not annotate the "no vaccination" edge with zero cost. Once you decide whether to vaccinate or not, there is a circle, indicating a random node (a random event; whether you get the disease or not) and, if you get it, another (minor or major)

Worked example 4.17 (Vaccination) It costs 10 to be vaccinated against a common disease. If you have the vaccination, the probability you will get the disease is $1e-7$. If you do not, the probability is 0.1. The disease is unpleasant; with probability 0.95, you will experience effects that cost you 1000 (eg several days in bed), but with probability 0.05, you will experience effects that cost you $1e6$. Should you be vaccinated?

Solution Figure 4.1 shows a decision tree for this problem. I have annotated some edges with the choices represented, and some edges with probabilities; the sum of probabilities over all rightward (downgoing) edges leaving a random node is 1. It is straightforward to compute expectations. The expected cost of the disease is $0.95 \times 1000 + 0.05 \times 1e6 = 50,950$. If you are vaccinated, your expected income will be $-(10 + 1e-7 \times 50,950) \approx -10.01$. If you are not, your expected income is $-5,095$. You should be vaccinated.

Example 4.17 has some subtleties. The conclusion is a rather shaky, though very common, use of the weak law. It's shaky, because the weak law has nothing to say about the outcome of a decision that you make only once. The proper interpretation of the example is that, if you had to make the choice many times over under the same set of circumstances, you should choose to be vaccinated. Notice you have to be careful using the example to argue that everyone should be vaccinated, because if lots of people were vaccinated then the probability of getting the disease would change. Since this probability goes down, the conclusion is fine, but you have to be careful about how you get there.

Sometimes there is more than one decision. We can still do simple examples, though drawing a decision tree is now quite important, because it allows us to keep track of cases and avoid missing anything. For example, assume I wish to buy a cupboard. Two nearby towns have used furniture shops (usually called antique shops these days). One is further away than the other. If I go to town A, I will have time to look in two (of three) shops; if I go to town B, I will have time to look in one (of two) shops. I could lay out this sequence of decisions (which town to go to; which shop to visit when I get there) as Fig. 4.2.

You should notice that this figure is missing a lot of information. What is the probability that I will find what I'm looking for in the shops? What is the value of finding it? What is the cost of going to each town? and so on. This information is not always easy to obtain. In fact, I might simply need to give my best subjective guess of these numbers. Furthermore, particularly if there are several decisions, computing the expected value of each possible sequence could get difficult. There are some kinds of model where one can compute expected values easily, but a good viable hypothesis about why people don't make optimal decisions is that optimal decisions are actually too hard to compute.

4.4.5 Utility

Sometimes it is hard to work with money. For example, in the case of a serious disease, choosing treatments often boils down to expected survival times, rather than money.

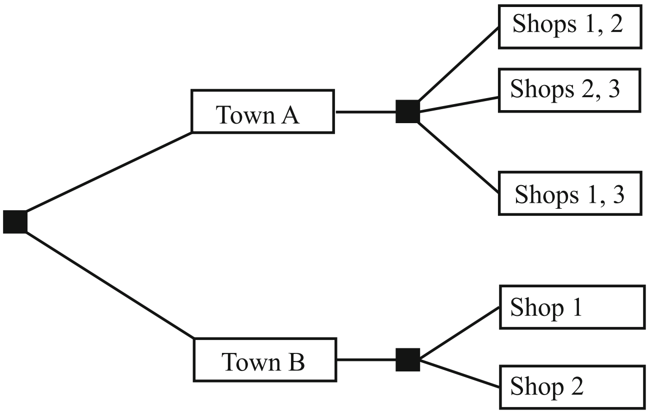

Fig. 4.2 The decision tree for the example of visiting furniture shops. Town A is nearer than town B, so if I go there I can choose to visit two of the three shops there; if I go to town B, I can visit only one of the two shops there. To decide what to do, I could fill in the probabilities and values of outcomes, compute the expected value of each pair of decisions, and choose the best. This could be tricky to do (where do I get the probabilities from?) but offers a rational and principled way to make the decision

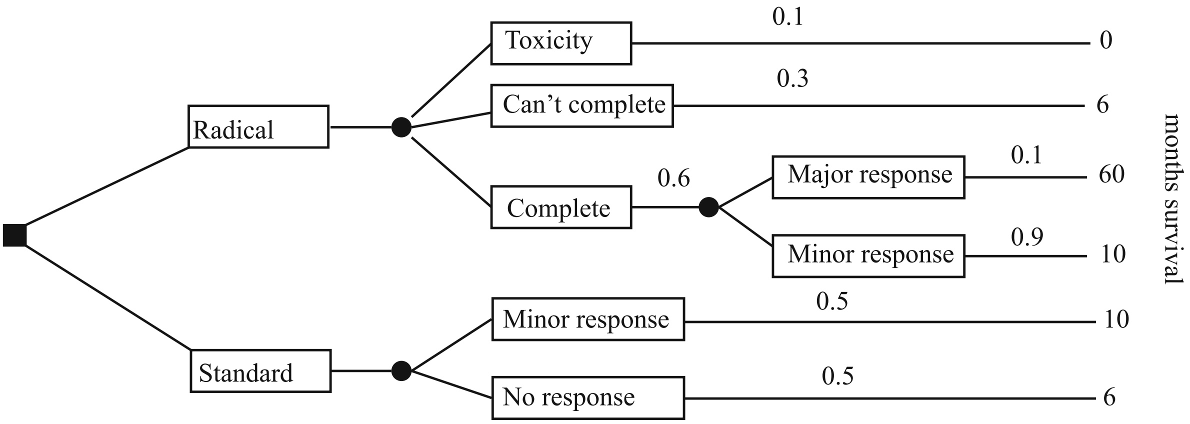

Fig. 4.3 A decision tree for Example 4.18

Worked example 4.18 (Radical Treatment) Imagine you have a nasty disease. There are two kinds of treatment: standard, and radical. Radical treatment might kill you (with probability 0.1); might be so damaging that doctors stop (with probability 0.3); but otherwise you will complete the treatment. If you do complete radical treatment, there could be a major response (probability 0.1) or a minor response. If you follow standard treatment, there could be a major response (probability 0.5) or a minor response, but the outcomes are less good. All this is best summarized in a decision tree (Fig. 4.3). What gives the longest expected survival time?

Solution In this case, expected survival time with radical treatment is $(0.1\times 0+0.3\times 6+0.6\times(0.1\times 60+0.9\times 10)) =$ 10.8 months; expected survival time without radical treatment is $0.5 \times 10 + 0.5 \times 6 = 8$ months.

Working with money values is not always a good idea. For example, many people play state lotteries. The expected value of a 1 bet on a state lottery is well below 1—why do people play? It's easy to assume that all players just can't do sums, but many players are well aware that the expected value of a bet is below the cost. It seems to be the case that people value money in a way that doesn't depend linearly on the amount of money. So, for example, people may value a million dollars rather more than a million times the value they place on one dollar. If this is true, we need some other way to keep track of value; this is sometimes called **utility**. It turns out to be quite hard to know how people value things, and there is quite good evidence that (a) human utility is complicated and (b) it is difficult to explain human decision making in terms of expected utility.

Worked example 4.19 (Human Utility is Not Expected Payoff) Here are four games:

- **Game 1:** The player is given 1. A biased coin is flipped, and the money is taken back with probability p; otherwise, the player keeps it.
- **Game 2:** The player stakes 1, and a fair coin is flipped; if the coin comes up heads, the player gets r and the stake back, but otherwise loses the original stake.
- **Game 3:** The player bets nothing; a biased coin is flipped, and if it comes up heads (probability q), the player gets $1e6$.
- **Game 4:** The player stakes 1000; a fair coin is flipped, and if it comes up heads, the player gets s and the stake back, but otherwise loses the original stake.

In particular, what happens if $r = 3 - 2p$ and $q = (1 - p)/1e6$ and $s = 2 - 2p + 1000$?

Solution Game 1 has expected value $(1 - p)1$. Game 2 has expected value $(1/2)(r - 1)$. Game 3 has expected value $q1e6$. Game 4 has expected value $(1/2)s - 500$.

In the case given, each game has the same expected value. Nonetheless, people usually have decided preferences for which game they would play. Generally, 4 is unattractive (seems expensive to play); 3 seems like free money, and so a good thing; 2 might be OK but is often seen as uninteresting; and 1 is unattractive. This should suggest to you that people's reasoning about money and utility is not what simple expectations predict.

4.5 You Should

4.5.1 Remember These Definitions

4.5.2 Remember These Terms

density . . . 91
normalizing . . . 92
normalizing constant . . . 92
standard deviation . . . 98
descriptive statistics . . . 98
empirical distribution . . . 99
trial . . . 99
sample . . . 99
realization . . . 99
independent identically distributed samples . . . 99
iid samples . . . 99
indicator function . . . 100
odds . . . 104
decision tree . . . 105
utility . . . 107

4.5.3 Use and Remember These Facts

Properties of probability density functions . . . 92
Expectations are linear . . . 95
Properties of variance . . . 96
Variance, a useful expression . . . 96
Covariance, useful expression . . . 97
Independent random variables have zero covariance . . . 97
Variance as covariance . . . 98

4.5.4 Remember These Points

4.5.5 Be Able to

- Interpret notation for joint and conditional probability for random variables; in particular, understand notation such as: $P(\{X\})$, $P(\{X = x\})$, $p(x)$, $p(x, y)$, $p(x|y)$
- Interpret a probability density function $p(x)$ as $P(\{X \in [x, x + dx]\})$.
- Interpret the expected value of a discrete random variable.
- Interpret the expected value of a continuous random variable.
- Compute expected values of random variables for straightforward cases.
- Write down expressions for mean, variance and covariance for random variables.
- Write out a decision tree.
- Exploit the weak law of large numbers.

Problems

Joint and Conditional Probability for Random Variables

4.1 A roulette wheel has one zero. Write X for the random variable representing the number that will come up on the wheel. What is the probability distribution of X?

4.2 Define a random variable X by the following procedure. Draw a card from a standard deck of playing cards. If the card is knave, queen, or king, then $X = 11$. If the card is an ace, then $X = 1$; otherwise, X is the number of the card (i.e. two through ten). Now define a second random variable Y by the following procedure. When you evaluate X, you look at the color of the card. If the card is red, then $Y = X - 1$; otherwise, $Y = X + 1$.

(a) What is $P(\{X \leq 2\})$?
(b) What is $P(\{X \geq 10\})$?
(c) What is $P(\{X \geq Y\})$?
(d) What is the probability distribution of $Y - X$?
(e) What is $P(\{Y \geq 12\}$?

4.3 Define a random variable by the following procedure. Flip a fair coin. If it comes up heads, the value is 1. If it comes up tails, roll a die: if the outcome is 2 or 3, the value of the random variable is 2. Otherwise, the value is 3.

(a) What is the probability distribution of this random variable?
(b) What is the cumulative distribution of this random variable?

4.4 Define three random variables, X, Y and Z by the following procedure. Roll a six-sided die and a four-sided die. Now flip a coin. If the coin comes up heads, then X takes the value of the six-sided die and Y takes the value of the four-sided die. Otherwise, X takes the value of the four-sided die and Y takes the value of the six-sided die. Z always takes the value of the sum of the dice.

(a) What is $P(X)$, the probability distribution of this random variable?
(b) What is $P(X, Y)$, the joint probability distribution of these two random variables?
(c) Are X and Y independent?
(d) Are X and Z independent?

4.5 Define two random variables X and Y by the following procedure. Flip a fair coin; if it comes up heads, then $X = 1$, otherwise $X = -1$. Now roll a six-sided die, and call the value U. We define $Y = U + X$.

(a) What is $P(Y|X = 1)$?
(b) What is $P(X|Y = 0)$?
(c) What is $P(X|Y = 7)$?
(d) What is $P(X|Y = 3)$?
(e) Are X and Y independent?

4.6 Magic the Gathering is a popular card game. Cards can be land cards, or other cards. We consider a game with two players. Each player has a deck of 40 cards. Each player shuffles their deck, then deals seven cards, called their *hand*. The rest of each player's deck is called their *library*. Assume that player one has 10 land cards in their deck and player two has 20. Write L_1 for the number of lands in player one's hand and L_2 for the number of lands in player two's hand. Write L_t for the number of lands in the top 10 cards of player one's library.

(a) Write $S = L_1 + L_2$. What is $P(\{S = 0\})$?
(b) Write $D = L_1 - L_2$. What is $P(\{D = 0\})$?
(c) What is the probability distribution for L_1?
(d) Write out the probability distribution for $P(L_1|L_t = 10)$.
(e) Write out the probability distribution $P(L_1|L_t = 5)$.

Continuous Random Variables

4.7 A continuous random variable has probability density function $p(x)$ which is proportional to $g(x)$, where

$$g(x) = \begin{cases} 0 & \text{if } x < -\frac{\pi}{2} \\ 0 & \text{if } x > \frac{\pi}{2} \\ \cos(x) & \text{otherwise} \end{cases} .$$

Write c for the constant of proportionality, so that $p(x) = cg(x)$.

(a) What is c? (you can look up the integral if you want)
(b) What is $P(\{X \geq 0\})$ (i.e. the probability you will observe a value greater than 0)? (you can look up the integral if you want)
(c) What is $P(\{|X| \leq 1\})$? (you can look up the integral if you want)

4.8 There is some (small!) voltage over the terminals of a warm resistor caused by noise (electrons moving around in the heat and banging into one another). This is a good example of a continuous random variable, and we can assume there is some probability density function for it, say $p(x)$. We assume that $p(x)$ has the property that

$$\lim_{\epsilon \to 0} \int_{v-\epsilon}^{v+\epsilon} p(x)dx = 0$$

which is what you'd expect for any function you're likely to have dealt with. Now imagine I define a new random variable by the following procedure: I flip a coin; if it comes up heads, I report 0; if tails, I report the voltage over the resistor. This random variable, u, has a probability 1/2 of taking the value 0, and 1/2 of taking a value from $p(x)$. Write this random variable's probability density function $q(u)$.

(a) Show that

$$\lim_{\epsilon \to 0} \int_{-\epsilon}^{\epsilon} q(u)du = \frac{1}{2}$$

(b) Explain why this is odd behavior.

Expected Values

4.9 Magic the Gathering is a popular card game. Cards can be land cards, or other cards. We consider a game with two players. Each player has a deck of 40 cards. Each player shuffles their deck, then deals seven cards, called their *hand*. The rest of each player's deck is called their *library*. Assume that player one has 10 land cards in their deck and player two has 20. Write L_1 for the number of lands in player one's hand and L_2 for the number of lands in player two's hand. Write L_t for the number of lands in the top 10 cards of player one's library.

(a) What is $\mathbb{E}[L_1]$?
(b) What is $\mathbb{E}[L_2]$?
(c) What is $\mathsf{var}[L_1]$?

4.10 A simple coin game is as follows: we have a box, which starts empty. P1 flips a fair coin. If it comes up heads, P2 gets the contents of the box, and the game ends. If it comes up tails, P1 puts a dollar in the box and they flip again; this repeats until it comes up heads

(a) With what probability will P2 win exactly 10 units?
(b) Write $S_\infty = \sum_{i=0}^{\infty} r^i$. Show that $(1 - r)S_\infty = 1$, so that

$$S_\infty = \frac{1}{1-r}$$

(c) Show that

$$\sum_{i=0}^{\infty} ir^i = (\sum_{i=1}^{\infty} r^i) + r(\sum_{i=1}^{\infty} r^i) + r^2(\sum_{i=1}^{\infty} r^i) + \dots$$

(look carefully at the limits of the sums!) and so show that

$$\sum_{i=0}^{\infty} ir^i = \frac{r}{(1-r)^2}.$$

(d) What is the expected value of the game? (you may find the results of the two previous subexercises helpful; they're not there just for show).

(e) How much should P2 pay to play, to make the game fair?

4.11 A simple card game is as follows. P1 pays a stake of 1 to play. P1 and P2 then each draw a card. If both cards are the same color, P2 keeps the stake and the game ends. If they are different colors, P2 pays P1 the stake and 1 extra (a total of 2).

(a) What is the expected value of the game to P1?

(b) P2 modifies the game, as follows. If both cards are court cards (that is, knave, queen, king), then P2 keeps the stake and the game ends; otherwise, the game works as before. Now what is the expected value of the game to P1?

4.12 A coin game that is occasionally played is "odd one out". In this game, there are rounds. In a round, each person flips a coin. There is an odd person out in that round if all but one have H and the other has T, OR all but one have T and the other has H.

(a) Three people play one round. What is the probability that there is an odd person out?

(b) Now four people play one round. What is the probability that there is an odd person out?

(c) Five people play until there is an odd person out. What is the expected number of rounds that they will play? (you can save yourself quite a lot of calculation by reading Sect. 5.1.3, if you don't mind skipping ahead a bit).

Mean, Variance and Covariance

4.13 Show that $\mathsf{var}[kX] = k^2\mathsf{var}[X]$.

4.14 Show that if X and Y are independent random variables, then $\mathsf{var}[X + Y] = \mathsf{var}[X] + \mathsf{var}[Y]$. You will find it helpful to remember that, for X and Y independent, $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.

Expectations and Descriptive Statistics

4.15 We have a dataset $\{x\}$ of N numbers, where the i'th number is x_i. Write X for the random variable that takes the i'th value with probability $1/N$, and every other value with zero probability; write $P(X)$ for the probability distribution of that random variable.

(a) Show that

$$\mathsf{mean}\,(\{x\}) = \mathbb{E}_{P(X)}[X].$$

(b) Show that

$$\mathsf{var}\,(\{x\}) = \mathsf{var}[X].$$

(c) Choose some function f. Write $\{f\}$ for the dataset whose i'th item is $f(x_i)$. Write F for the random variable $f(X)$. Show that

$$\mathsf{mean}\,(\{f\}) = \mathbb{E}[F] = \mathbb{E}_{P(X)}[f].$$

Markov and Chebyshev Inequalities

4.16 The random variable X takes the values -2, -1, 0, 1, 2, but has an unknown probability distribution. You know that $\mathbb{E}[\|X\|] = 0.2$. Use Markov's inequality to give a *lower* bound on $P(\{X = 0\})$. *Hint:* Notice that $P(\{X = 0\}) = 1 - P(\{\|X\| = 1\}) - P(\{\|X\|\} = 2)$.

4.17 The random variable X takes the values 1, 2, 3, 4, 5, but has unknown probability distribution. You know that $\mathbb{E}[X] = 2$ and $\mathsf{var}(\{X\}) = 0.01$. Use Chebyshev's inequality to give a *lower* bound on $P(\{X = 2\})$.

4.18 You have a biased random number generator. This generator produces a random number with mean value -1, and standard deviation 0.5. Write $\mathcal{A}$ for the event that the number generator produces a non-negative number. Use Chebyshev's inequality to bound $P(\mathcal{A})$.

4.19 You observe a random number generator. You know that it can produce the values $-2, -1$, 0, 1, or 2. You are told that it has been adjusted so that: (1) the mean value it produces is zero and; (2) the standard deviation of the numbers it produces is 1.

(a) Write $\mathcal{A}$ for the event that the number generator produces a number that is not 0. Use Chebyshev's inequality to bound $P(\mathcal{A})$.
(b) Write $\mathcal{B}$ for the event that the number generator produces -2 or 2. Use Chebyshev's inequality to bound $P(\mathcal{B})$.

Using Expectations

4.20 Two players P1 and P2 agree to play the following game. Each puts up a stake of 1 unit. They will play seven rounds, where each round involves flipping a fair coin. If the coin comes up H, P1 wins the round, otherwise P2 wins. The first player to win four rounds gets both stakes. After four rounds, P1 has won three rounds and P2 has won one round, but they have to stop. What is the fairest way to divide the stakes?

4.21 Imagine we have a game with two players, who are playing for a stake. There are no draws, the winner gets the whole stake, and the loser gets nothing. The game must end early. We decide to give each player the expected value of the game for that player, from that state. Show that the expected values add up to the value of the stake (i.e. there won't be too little or too much money in the stake.

Programming Exercises

4.22 An airline company runs a flight that has six seats. Each passenger who buys a ticket has a probability p of turning up for the flight. These events are independent.

(a) The airline sells six tickets. What is the expected number of passengers, if $p = 0.9$?
(b) How many tickets should the airline sell to ensure that the expected number of passengers is greater than six, if $p = 0.7$? **Hint:** The easiest way to do this is to write a quick program that computes the expected value of passengers that turn up for each the number of tickets sold, then search the number of tickets sold.

4.23 An airline company runs a flight that has 10 seats. Each passenger who buys a ticket has a probability p of turning up for the flight. The gender of the passengers is not known until they turn up for a flight, and women buy tickets with the same frequency that men do. The pilot is eccentric, and will not fly unless at least two women turn up.

(a) How many tickets should the airline sell to ensure that the expected number of passengers that turn up is greater than 10?
(b) The airline sells 10 tickets. What is the expected number of passengers on the aircraft, given that it flies? (i.e. that at least two women turn up). Estimate this value with a simulation.

4.24 We will investigate the weak law of large numbers using simulations. Write X for a random variable that takes the values -1 and 1 with equal probability, and no other value. Clearly, $\mathbb{E}[X] = 0$. Write $X^{(N)}$ for the random variable obtained by drawing N samples of X, then averaging them.

(a) For each N in $\{1, 10, 20, \ldots, 100\}$, simulate 1000 samples of $X^{(N)}$. Produce a graph showing a boxplot of these samples for each N, plotted against N. What do you notice?

(b) For each N in $\{1, 10, 20, \ldots, 100\}$, simulate 1000 samples of $X^{(N)}$. Produce a graph showing the variance of these samples as a function of $1/N$. What do you notice?

(c) Show that the normal approximation of a binomial distribution suggests that about 68% of the observed values of $X^{(N)}$ lie in the range

$$\left[-\frac{1}{2\sqrt{N}}, \frac{1}{2\sqrt{N}}\right].$$

(d) For each N in $\{1, 10, 20, \ldots, 100\}$, simulate 1000 samples of $X^{(N)}$. For each N, compute the 84% quantile ($q_{84\%}$) and the 16% quantile ($q_{16\%}$). Now compute

$$\alpha = \max(|q_{84\%}|, |q_{16\%}|).$$

This α should have the property that about 68% of the observed values lie in the range $[-\alpha, \alpha]$. Now plot $1/\alpha^2$ as a function of N. What do you notice?