Random Variables and Expectation

Question: The homeworks of 20 students are collected in, randomly shuffled and returned to the students. How many students receive their own homework?

To answer this question, we first need to specify the probability space: plainly, it should consist of all 20! permutations of the homeworks, each with probability $\frac{1}{20!}$. [Note that this is the same as the probability space for card shuffling, except that the number of items being shuffled is now 20 rather than 52.] It helps to have a picture of a permutation. Think of 20 books lined up on a shelf, labeled from left to right with $1,2,\ldots,20$. A permutation π is just a reordering of the books, which we can describe just by listing their labels from left to right. Let's denote by π_i the label of the book that is in position i. We are interested in the number of books that are still in their original position, i.e., in the number of i's such that $\pi_i = i$. These are often known as fixed points of the permutation.

Of course, our question does not have a simple numerical answer (such as 6), because the number depends on the particular permutation we choose (i.e., on the sample point). Let's call the number of fixed points X. To make life simpler, let's also shrink the class size down to 3 for a while. The following table gives a complete listing of the sample space (of size $3! = 6$), together with the corresponding value of X for each sample point. [We use our bookshelf convention for writing a permutation: thus, for example, the permutation 312 means that book 3 is on the left, book 1 in the center, and book 2 on the right. You should check you agree with this table.]

permutation π	value of X
123	3
132	1
213	1
231	0
312	0
321	1

Thus we see that X takes on values 0, 1 or 3, depending on the sample point. A quantity like this, which takes on some numerical value at each sample point, is called a *random variable* (or *r.v.*) on the sample space.

Definition 17.1 (random variable): A random variable X on a sample space Ω is a function that assigns to each sample point $\omega \in \Omega$ a real number $X(\omega)$.

Until further notice, we'll restrict our attention to discrete random variables: i.e., their values will be integers or rationals, rather than arbitrary real numbers.

The r.v. X in our permutation example above is completely specified by its values at all sample points, as given in the above table. (Thus, for example, $X(123) = 3$ etc.)

Rather than the value at each sample point, we are usually more interested in the *set* of points at which the r.v. takes on some given value. Let a be any real number. Then the set

$$\{\omega \in \Omega : X(\omega) = a\}$$

is an *event* in the sample space (why?). We usually abbreviate this event to simply "$X = a$". Since $X = a$ is an event, we can talk about its probability, $\Pr[X = a]$. The collection of these probabilities, for all possible values of a, is known as the *distribution* of the r.v. X.

Definition 17.2 (distribution): The distribution of a discrete random variable X is the collection of values $\{(a, \Pr[X = a]) : a \in \mathscr{A}\}$, where $\mathscr{A}$ is the set of all possible values taken by X. [In most of our examples, $\mathscr{A}$ will be (some subset of) the integers.]

Thus the distribution of the random variable X in our permutation example above is

$$\Pr[X = 0] = \frac{1}{3}; \qquad \Pr[X = 1] = \frac{1}{2}; \qquad \Pr[X = 3] = \frac{1}{6};$$

and $\Pr[X = a] = 0$ for all other values of a.

Notice that the sum of the probabilities $\Pr[X = a]$ over all possible values of a is exactly 1. *This must always be the case, for any random variable.* Why? Because any random variable X must take on one, and only one, value $X(\omega)$ at every sample point ω. I.e., the events $X = a$ *partition* the sample space. So when we sum up the probabilities of the events $X = a$, we are really summing up the probabilities of all the sample points.

Expectation

In most applications, the complete distribution of a r.v. is very hard to calculate: for example, suppose we go back to our original question with 20 students. In principle, we'd have to enumerate $20! \approx 2.4 \times 10^{18}$ sample points, compute the value of X at each one, and count the number of points at which X takes on each of its possible values! Moreover, even when we can compute the complete distribution of a r.v., it is often not very informative.

For these reasons, we seek to *compress* the distribution into a more compact, convenient form that is also easier to compute. The most widely used such form is the *expectation* (or *mean* or *average*) of the r.v.

Definition 17.3 (expectation): The expectation of a discrete random variable X is defined as

$$\mathrm{E}(X) = \sum_{a \in \mathscr{A}} a \times \Pr[X = a],$$

where the sum is over all possible values taken by the r.v.

For our running permutation example, the expectation is

$$\mathrm{E}(X) = \left(0 \times \frac{1}{3}\right) + \left(1 \times \frac{1}{2}\right) + \left(3 \times \frac{1}{6}\right) = 0 + \frac{1}{2} + \frac{1}{2} = 1.$$

I.e., the expected number of fixed points in a permutation of three items is exactly 1.

The expectation can be seen in some sense as a "typical" value of the r.v. (though note that it may not actually be a value that the r.v. ever takes on). The question of how typical the expectation is for a given r.v. is a very important one that we shall return to in a later lecture.

Here are some simple examples of expectations.

1. **Single die.** Throw one fair die. Let X be the number that comes up. Then X takes on values $1,2,\ldots,6$ each with probability $\frac{1}{6}$, so

$$\mathrm{E}(X) = \frac{1}{6}(1+2+3+4+5+6) = \frac{21}{6} = \frac{7}{2}.$$

Note that X never actually takes on its expected value $\frac{7}{2}$.

2. **Two dice.** Throw two fair dice. Let X be the sum of their scores. Then the distribution of X is

a	2	3	4	5	6	7	8	9	10	11	12
$\Pr[X=a]$	$\frac{1}{36}$	$\frac{1}{18}$	$\frac{1}{12}$	$\frac{1}{9}$	$\frac{5}{36}$	$\frac{1}{6}$	$\frac{5}{36}$	$\frac{1}{9}$	$\frac{1}{12}$	$\frac{1}{18}$	$\frac{1}{36}$

The expectation is therefore

$$\mathrm{E}(X) = \left(2\times\frac{1}{36}\right) + \left(3\times\frac{1}{18}\right) + \left(4\times\frac{1}{12}\right) + \cdots + \left(12\times\frac{1}{36}\right) = 7.$$

3. **Roulette.** A roulette wheel is spun. You bet \$1 on Black. If a black number comes up, you receive your stake plus \$1; otherwise you lose your stake. Let X be your net winnings in one game. Then X can take on the values $+1$ and -1, and $\Pr[X=1] = \frac{18}{38}$, $\Pr[X=-1] = \frac{20}{38}$. [Recall that a roulette wheel has 38 slots: the numbers $1,2,\ldots,36$, half of which are red and half black, plus 0 and 00, which are green.] Thus

$$\mathrm{E}(X) = \left(1\times\frac{18}{38}\right) + \left(-1\times\frac{20}{38}\right) = -\frac{1}{19};$$

i.e., you expect to lose about a nickel per game. Notice how the zeros tip the balance in favor of the casino!

Linearity of expectation

So far, we've computed expectations by brute force: i.e., we have written down the whole distribution and then added up the contributions for all possible values of the r.v. The real power of expectations is that in many real-life examples they can be computed much more easily using a simple shortcut. The shortcut is the following:

Theorem 17.1: *For any two random variables X and Y on the same probability space, we have*

$$\mathrm{E}(X+Y) = \mathrm{E}(X) + \mathrm{E}(Y).$$

Also, for any constant c, we have

$$\mathrm{E}(cX) = c\mathrm{E}(X).$$

Proof: To make the proof easier, we will first rewrite the definition of expectation in a more convenient form. Recall from Definition 18.3 that

$$\mathrm{E}(X) = \sum_{a\in\mathscr{A}} a\times\Pr[X=a].$$

Let's look at one term $a\times\Pr[X=a]$ in the above sum. Notice that $\Pr[X=a]$, by definition, is the sum of $\Pr[\omega]$ over those sample points ω for which $X(\omega)=a$. And we know that every sample point $\omega\in\Omega$ is in exactly one of these events $X = a$. This means we can write out the above definition in a more long-winded form as

$$\mathrm{E}(X) = \sum_{\omega\in\Omega} X(\omega) \times \Pr[\omega]. \tag{1}$$

This equivalent definition of expectation will make the present proof much easier (though it is usually less convenient for actual calculations).

Now let's write out $\mathrm{E}(X+Y)$ using equation (1).

$$\begin{aligned}
\mathrm{E}(X+Y) &= \textstyle\sum_{\omega\in\Omega}(X+Y)(\omega) \times \Pr[\omega] \\
&= \textstyle\sum_{\omega\in\Omega}(X(\omega)+Y(\omega)) \times \Pr[\omega] \\
&= \textstyle\sum_{\omega\in\Omega}\Big(X(\omega) \times \Pr[\omega]\Big) + \sum_{\omega\in\Omega}\Big(Y(\omega) \times \Pr[\omega]\Big) \\
&= \mathrm{E}(X) + \mathrm{E}(Y).
\end{aligned}$$

In the last step, we used equation (1) twice.

This completes the proof of the first equality. The proof of the second equality is much simpler and is left as an exercise. □

Theorem 18.1 is very powerful: it says that the expectation of a sum of r.v.'s is the sum of their expectations, with no assumptions about the r.v.'s. (Note that we didn't require, for example, that X and Y are independent.) We can use Theorem 18.1 to conclude things like $\mathrm{E}(3X-5Y) = 3\mathrm{E}(X) - 5\mathrm{E}(Y)$. This property is known as linearity of expectation. *Important caveat: Theorem 18.1 does* **not** *say that* $\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y)$, *or that* $\mathrm{E}(\frac{1}{X}) = \frac{1}{\mathrm{E}(X)}$ *etc. These claims are not true in general. It is only sums and differences and constant multiples of random variables that behave so nicely.*

Now let's see some examples of Theorem 18.1 in action.

4. **Two dice again.** Here's a much less painful way of computing $\mathrm{E}(X)$, where X is the sum of the scores of the two dice. Note that $X = X_1 + X_2$, where X_i is the score on die i. We know from example 1 above that $\mathrm{E}(X_1) = \mathrm{E}(X_2) = \frac{7}{2}$. So by Theorem 18.1 we have $\mathrm{E}(X) = \mathrm{E}(X_1) + \mathrm{E}(X_2) = 7$.

5. **More roulette.** Suppose we play the above roulette game not once, but 1000 times. Let X be our expected net winnings. Then $X = X_1 + X_2 + \ldots + X_{1000}$, where X_i is our net winnings in the ith play. We know from earlier that $\mathrm{E}(X_i) = -\frac{1}{19}$ for each i. Therefore, by Theorem 18.1, $\mathrm{E}(X) = \mathrm{E}(X_1) + \mathrm{E}(X_2) + \cdots + \mathrm{E}(X_{1000}) = 1000 \times (-\frac{1}{19}) = -\frac{1000}{19} \approx -53$. So if you play 1000 games, you expect to lose about \$53.

6. **Homeworks.** Let's go back and answer our original question about the class of 20 students. Recall that the r.v. X is the number of students who receive their own homework after shuffling (or equivalently, the number of fixed points). To take advantage of Theorem 18.1, we need to write X as the *sum* of simpler r.v.'s. But since X *counts* the number of times something happens, we can write it as a sum using the following trick:

$$X = X_1 + X_2 + \ldots + X_{20}, \qquad \text{where } X_i = \begin{cases} 1 & \text{if student } i \text{ gets her own hw;} \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

[You should think about this equation for a moment. Remember that all the X's are random variables. What does an equation involving random variables mean? What we mean is that, *at every sample point* ω, we have $X(\omega) = X_1(\omega) + X_2(\omega) + \ldots + X_{20}(\omega)$. Why is this true?]

A 0/1-valued random variable such as X_i is called an indicator random variable of the corresponding event (in this case, the event that student i gets her own hw). For indicator r.v.'s, the expectation is particularly easy to calculate. Namely,

$$E(X_i) = (0 \times \Pr[X_i = 0]) + (1 \times \Pr[X_i = 1]) = \Pr[X_i = 1].$$

But in our case, we have

$$\Pr[X_i = 1] = \Pr[\text{student } i \text{ gets her own hw}] = \frac{1}{20}.$$

Now we can apply Theorem 18.1 to (2), to get

$$E(X) = E(X_1) + E(X_2) + \cdots + E(X_{20}) = 20 \times \frac{1}{20} = 1.$$

So we see that the expected number of students who get their own homeworks in a class of size 20 is 1. But this is exactly the same answer as we got for a class of size 3! And indeed, we can easily see from the above calculation that we would get $E(X) = 1$ for *any* class size n: this is because we can write $X = X_1 + X_2 + \ldots + X_n$, and $E(X_i) = \frac{1}{n}$ for each i.

So the expected number of fixed points in a random permutation of n items is always 1, regardless of n. Amazing, but true.

7. **Coin tosses.** Toss a fair coin 100 times. Let the r.v. X be the number of Heads. As in the previous example, to take advantage of Theorem 18.1 we write

$$X = X_1 + X_2 + \ldots + X_{100},$$

where X_i is the indicator r.v. of the event that the ith toss is Heads. Since the coin is fair, we have

$$E(X_i) = \Pr[X_i = 1] = \Pr[i\text{th toss is Heads}] = \tfrac{1}{2}.$$

Using Theorem 18.1, we therefore get

$$E(X) = \sum_{i=1}^{100} = 100 \times \tfrac{1}{2} = 50.$$

More generally, the expected number of Heads in n tosses of a fair coin is $\frac{n}{2}$. And in n tosses of a biased coin with Heads probability p, it is np (why?).

8. **Balls and bins.** Throw m balls into n bins. Let the r.v. X be the number of balls that land in the first bin. Then X behaves exactly like the number of Heads in n tosses of a biased coin, with Heads probability $\frac{1}{n}$ (why?). So from example 7 we get $E(X) = \frac{m}{n}$.

In the special case $m = n$, the expected number of balls in any bin is 1. If we wanted to compute this directly from the distribution of X, we'd get into a messy calculation involving binomial coefficients, a bit like the load balancing example in the previous lecture.

Here's another example on the same sample space. Let the r.v. Y be the number of empty bins. The distribution of Y is horrible to contemplate: to get a feel for this, you might like to write it down for $m = n = 3$ (3 balls, 3 bins). However, computing the expectation $E(Y)$ is a piece of cake using Theorem 18.1. As usual, let's write

$$Y = Y_1 + Y_2 + \cdots + Y_n, \tag{3}$$

where Y_i is the indicator r.v. of the event "bin i is empty". Again as usual, the expectation of Y_i is easy:

$$\mathrm{E}(Y_i) = \Pr[Y_i = 1] = \Pr[\text{bin } i \text{ is empty}] = \left(1 - \frac{1}{n}\right)^m;$$

recall that we computed this probability (quite easily) way back in lecture 17. Applying Theorem 18.1 to (3) we therefore have

$$\mathrm{E}(Y) = \sum_{i=1}^{n} \mathrm{E}(Y_i) = n\left(1 - \frac{1}{n}\right)^m,$$

a very simple formula, very easily derived.

Let's see how it behaves in the special case $m = n$ (same number of balls as bins). In this case we get $\mathrm{E}(Y) = n(1 - \frac{1}{n})^n$. Now the quantity $(1 - \frac{1}{n})^n$ can be approximated (for large enough values of n) by the number $\frac{1}{\mathrm{e}}$.[1] So we see that

$$\mathrm{E}(Y) \to \frac{n}{\mathrm{e}} \approx 0.368n \quad \text{as } n \to \infty.$$

The bottom line is that, if we throw (say) 1000 balls into 1000 bins, the expected number of empty bins is about 368.

[1]More generally, it is a standard fact that for any constant c,

$$(1 + \tfrac{c}{n})^n \to \mathrm{e}^c \quad \text{as } n \to \infty.$$

We just used this fact in the special case $c = -1$. The approximation is actually very good even for quite small values of n — try it yourself! E.g., for $n = 20$ we already get $(1 - \frac{1}{n})^n$=0.358, which is very close to $\frac{1}{\mathrm{e}} = 0.367\ldots$. The approximation gets better and better for larger n.