Random Variables: Distribution and Expectation

Recall our setup of a probabilistic experiment as a procedure of drawing a sample from a set of possible values, and assigning a probability for each possible outcome of the experiment. For example, if we toss a fair coin n times, then there are 2^n possible outcomes, each of which is equally likely and has probability $\frac{1}{2^n}$.

Now suppose we want to make a measurement in our experiment. For example, we can ask what is the number of heads in n coin tosses; call this number X. Of course, X is not a fixed number, but it depends on the actual sequence of coin flips that we obtain. For example, if $n = 4$ and we observe the outcome $\omega = HTHH$, then $X = 3$; whereas if we observe the outcome $\omega = HTHT$, then $X = 2$. In this example of n coin tosses, we only know that X is an integer between 0 and n, but we do not know what its exact value is until we observe which outcome of n coin flips is realized and count how many heads there are. Because every possible outcome is assigned a probability, the value X also carries with it a probability for each possible value it can take. The table below lists all the possible values X can take in the example of $n = 4$ coin tosses, along with their respective probabilities.

outcomes ω	value of X (# heads)	probability of occurring
$TTTT$	0	$1/16$
$HTTT$, $THTT$, $TTHT$, $TTTH$	1	$4/16$
$HHTT$, $HTHT$, $HTTH$, $THHT$, $THTH$, $TTHH$	2	$6/16$
$HHHT$, $HHTH$, $HTHH$, $THHH$	3	$4/16$
$HHHH$	4	$1/16$

Such a value X that depends on the outcome of the probabilistic experiment is called a *random variable* (abbreviated *r.v.*). As we see from the example above, a random variable X typically does not have a definitive value, but instead only has a probability *distribution* over the set of possible values X can take, which is why it is called random. So the question "What is the number of heads in n coin tosses?" does not exactly make sense because the answer X is a random variable. But the question "What is the *typical* number of heads in n coin tosses?" makes sense — it is asking what is the average value of X (the number of heads) if we repeat the experiment of tossing n coins multiple times. This average value is called the *expectation* of X, and is one of the most useful summaries (also called *statistics*) of an experiment.

1 Random Variables

Before we formalize the above notions, let us consider another example to enforce our conceptual understanding of a random variable.

Example: Fixed Points of Permutations

Question: Suppose we collect the homeworks of n students, randomly shuffle them, and return them to the students. How many students receive their own homework?

Here the probability space consists of all $n!$ permutations of the homeworks, each with equal probability $\frac{1}{n!}$. If we label the homeworks as $1, 2, \ldots, n$, then each sample point is a permutation $\pi = (\pi_1, \ldots, \pi_n)$ where π_i is the homework that is returned to the ith student. We call i a *fixed point* of π if $\pi_i = i$, i.e., if student i receives their own homework.

As in the coin flipping case above, our question does not have a simple numerical answer (such as 4), because the number depends on the particular permutation we choose (i.e., on the sample point). Let us call the number of fixed points X_n, which is a random variable taking values in the set $\{0, 1, 2, \ldots, n\}$. (Actually the value $X_n = n - 1$ is not possible: why?)

Formal Definition of a Random Variable

We now formalize the concepts discussed above.

Definition 16.1 (Random Variable)**.** *A random variable X on a sample space Ω is a function $X : \Omega \to \mathbb{R}$ that assigns to each sample point $\omega \in \Omega$ a real number $X(\omega)$.*

Until further notice, we will restrict our attention to random variables that are discrete, i.e., they take values in a range that is finite or countably infinite. This means even though we define X to map Ω to $\mathbb{R}$, the actual set of values $\{X(\omega) : \omega \in \Omega\}$ that X takes is a discrete subset of $\mathbb{R}$.

A random variable can be visualized in general by the picture in Figure 1.[1] Note that the term "random variable" is really something of a misnomer: it is a function so there is nothing random about it and it is definitely not a variable! What is random is which sample point of the experiment is realized and hence the value that the random variable maps the sample point to.

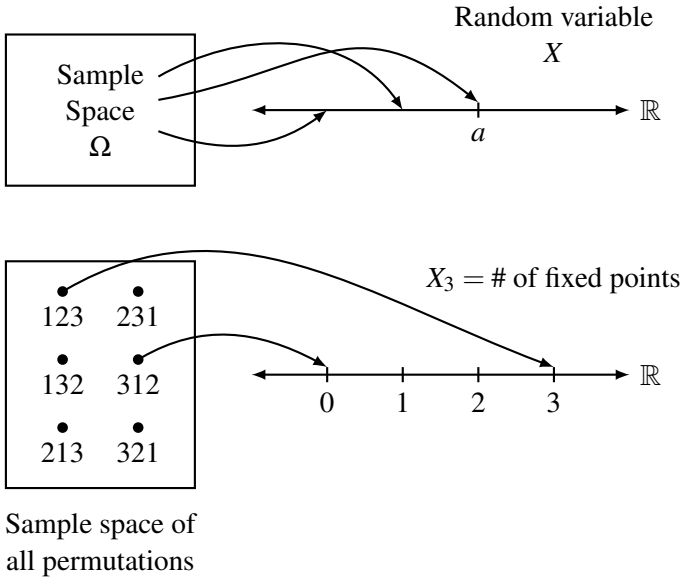

Figure 1: Visualization of how a random variable is defined on the sample space.

Exercise. By completing the lower picture in Figure 1, show that the only possible values for the r.v. X_3 are 0, 1 and 3, and that their probabilities are $\frac{1}{3}$, $\frac{1}{2}$ and $\frac{1}{6}$, respectively.

[1] The figures in this note are inspired by figures in Chapter 2 of *Introduction to Probability* by D. Bertsekas and J. Tsitsiklis.

2 Probability Distribution

When we introduced the basic probability space in an earlier note, we defined two things:

1. The sample space Ω consisting of all the possible outcomes (sample points) of the experiment.
2. The probability of each of the sample points.

Analogously, there are two important things about any random variable:

1. The set of values that it can take.
2. The probabilities with which it takes on the values.

Since a random variable is defined on a probability space, we can calculate these probabilities given the probabilities of the sample points. Let a be any number in the range of a random variable X. Then the set

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

is an *event* in the sample space (because it is a subset of Ω). We usually abbreviate this event to simply "$X = a$". Since $X = a$ is an event, we can talk about its probability, $\mathbb{P}[X = a]$. The collection of these probabilities, for all possible values of a, is known as the *distribution* of the random variable X.

Definition 16.2 (Distribution)**.** *The distribution of a discrete random variable X is the collection of values $\{(a, \mathbb{P}[X = a]) : a \in \mathscr{A}\}$, where $\mathscr{A}$ is the set of all possible values taken by X.*

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

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

If needed, we may also write $\mathbb{P}[X = a] = 0$ for all other values of a.

The distribution of a random variable can be visualized as a bar diagram, shown in Figure 2. The x-axis represents the values that the random variable can assume. The height of the bar at a value a is the probability $\mathbb{P}[X = a]$. Each of these probabilities can be computed by looking at the probability of the corresponding event in the sample space.

Note that the collection of events $X = a$, for $a \in \mathscr{A}$, satisfy two important properties:

- Any two events $X = a_1$ and $X = a_2$ with $a_1 \neq a_2$ are disjoint.
- The union of all these events is equal to the entire sample space Ω.

The collection of events thus form a *partition* of the sample space (see Figure 2). Both properties follow directly from the fact that X is a function defined on Ω, i.e., X assigns a unique value to each and every possible sample point in Ω. So, when we sum up the probabilities of the events $X = a$, we are really summing up the probabilities of all the sample points, giving us a total of exactly 1.

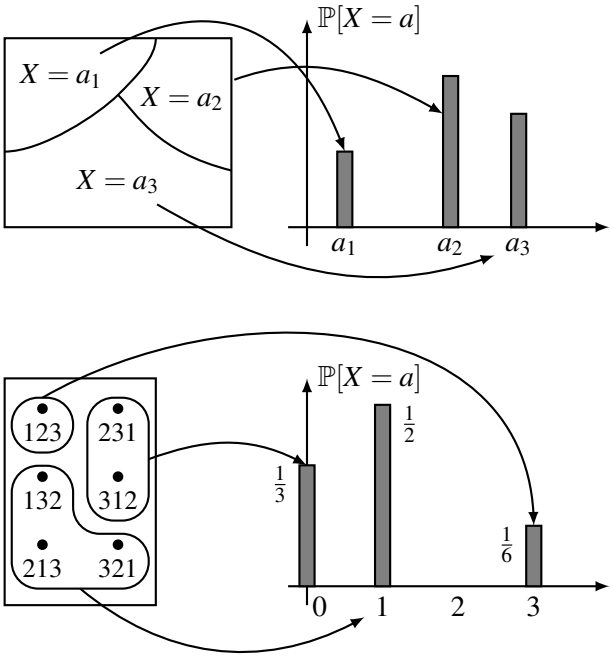

Figure 2: Visualization of how the distribution of a random variable is defined.

Bernoulli Distribution

A simple yet very useful probability distribution is the *Bernoulli* distribution of a random variable which takes value in $\{0,1\}$:

$$\mathbb{P}[X = i] = \begin{cases} p, & \text{if } i = 1, \\ 1-p, & \text{if } i = 0, \end{cases}$$

where $0 \leq p \leq 1$. We say that X is distributed as a *Bernoulli* random variable with parameter p, and write

$$X \sim \text{Bernoulli}(p) \qquad \text{or} \qquad X \sim \text{Ber}(p).$$

Binomial Distribution

Let us return to our coin tossing example above, where we defined our random variable X to be the number of heads. More formally, consider the random experiment consisting of n independent tosses of a biased coin that shows H with probability p. Each sample point ω is a sequence of tosses, and $X(\omega)$ is defined to be the number of heads in ω. For example, when $n = 3$, $X(THH) = 2$.

To compute the distribution of X, we first enumerate the possible values that X can take. They are simply $0, 1, \ldots, n$. Then we compute the probability of each event $X = i$ for $i = 0, 1, \ldots, n$. The probability of the event $X = i$ is the sum of the probabilities of all the sample points with exactly i heads (for example, if $n = 3$ and $i = 2$, there would be three such sample points $\{HHT, HTH, THH\}$). Any such sample point has probability $p^i(1-p)^{n-i}$, since the coin flips are independent. There are exactly $\binom{n}{i}$ of these sample points. Hence,

$$\mathbb{P}[X = i] = \binom{n}{i} p^i (1-p)^{n-i}, \qquad \text{for } i = 0, 1, \ldots, n. \tag{1}$$

This distribution, called the *binomial* distribution, is one of the most important distributions in probability. A random variable with this distribution is called a *binomial* random variable, and we write

$$X \sim \text{Bin}(n, p),$$

where n denotes the number of trials and p the probability of success (observing an H in the example). An example of a binomial distribution is shown in Figure 3. Notice that due to the properties of X mentioned above, it must be the case that $\sum_{i=0}^{n} \mathbb{P}[X = i] = 1$, which implies that $\sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} = 1$. This provides a probabilistic proof of the Binomial Theorem from an earlier note where we saw it combinatorially, for $a = p$ and $b = 1 - p$.

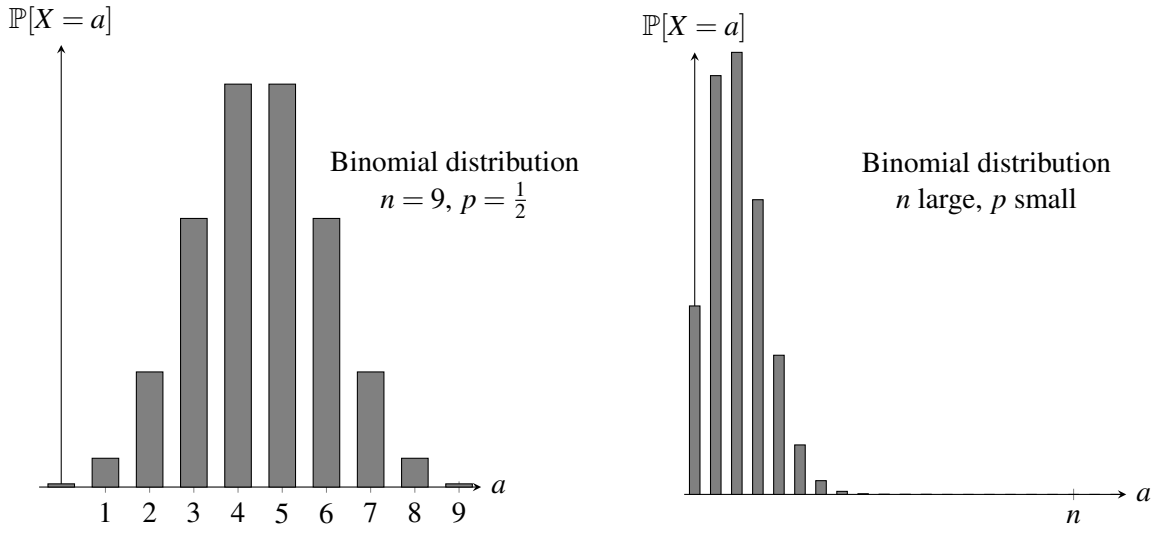

Figure 3: The binomial distributions for two choices of (n, p).

Although we define the binomial distribution in terms of an experiment involving tossing coins, this distribution is useful for modeling many real-world problems. Consider for example the error correction problem studied earlier. Recall that we wanted to encode n packets into $n+k$ packets such that the recipient can reconstruct the original n packets from any n packets received. But in practice, the number of packet losses is random, so how do we choose k, the amount of redundancy? If we model each packet getting lost with probability p and the losses are independent, then if we transmit $n+k$ packets, the number of packets received is a random variable X with binomial distribution: $X \sim \text{Bin}(n+k, 1-p)$ (we are tossing a coin $n+k$ times, and each coin turns out to be a head (packet received) with probability $1-p$). So the probability of successfully decoding the original data is:

$$\mathbb{P}[X \geq n] = \sum_{i=n}^{n+k} \mathbb{P}[X = i] = \sum_{i=n}^{n+k} \binom{n+k}{i} (1-p)^i p^{n+k-i}.$$

Given fixed n and p, we can choose k such that this probability is no less than, say, 0.99.

Hypergeometric Distribution

Consider an urn containing $N = B + W$ balls, where B balls are black and W are white. Suppose you randomly sample $n \leq N$ balls from the urn *with* replacement, and let X denote the number of black balls in your sample. What is the probability distribution of X? Since the probability of seeing a black ball is

B/N for each draw, independently of all other draws, X follows the binomial distribution $\text{Bin}(n,p)$, where $p = B/N$.

What if you randomly sample $n \leq N$ balls from the urn *without* replacement? In this case, the probability of seeing a black ball in the ith draw depends on the colors of the $i-1$ balls already drawn; that is, unlike in the case of sampling with replacement, the draws are not independent. The probability distribution of the number Y of black balls in this setting can be found as follows.

Consider a sequence ω of n draws where the first k balls are black and the next $n-k$ balls are white. Then, using the Product Rule from Note 14, the probability of this particular sequence of draws can be computed as

$$\mathbb{P}[\omega] = \frac{B}{N} \times \frac{B-1}{N-1} \times \cdots \times \frac{B-k+1}{N-k+1} \times \frac{W}{N-k} \times \frac{W-1}{N-k-1} \times \cdots \times \frac{W-(n-k)+1}{N-n+1} = \frac{\binom{B}{k}\binom{W}{n-k}}{\binom{N}{n}} \frac{1}{\binom{n}{k}}. \quad (2)$$

Furthermore, any other sequence ω' consisting of k blacks balls and $n-k$ white balls has exactly the same probability as ω: if we carry out a similar calculation as in (2) for the sequence of draws in ω', the numerators appearing in $\mathbb{P}[\omega']$ will be some permutation of the numerators in the first line of (2), while the denominators will be exactly the same as in (2). Since there are $\binom{n}{k}$ distinct sequences of length n consisting of k black balls and $n-k$ white balls, we obtain

$$\mathbb{P}[Y=k] = \binom{n}{k}\mathbb{P}[\omega] = \frac{\binom{B}{k}\binom{N-B}{n-k}}{\binom{N}{n}}, \quad (3)$$

for $k \in \{0,1,\ldots,n\}$. (Note that $\binom{m}{j} = 0$ if $j > m$, so $\mathbb{P}[Y=k] \neq 0$ only if $\max(0, n+B-N) \leq k \leq \min(n,B)$.)

This probability distribution is called the *hypergeometric distribution* with parameters N, B, n, and we write

$$Y \sim \text{Hypergeometric}(N,B,n).$$

3 Multiple Random Variables, Independence, and Exchangeability

Often one is interested in multiple random variables on the same sample space. Consider, for example, the sample space of flipping three coins. One could define many random variables: for example a random variable X indicating the number of heads, or a random variable Y indicating the number of tails, or a binary random variable Z indicating whether the first toss is H or not. Note that for each sample point, any random variable has a specific value: e.g., for $\omega = HTT$, we have $X(\omega) = 1$, $Y(\omega) = 2$, and $Z(\omega) = 1$.

The concept of a distribution can then be extended to probabilities for the combination of values for multiple random variables.

Definition 16.3. *The joint distribution of two discrete random variables X and Y is the collection of values $\{((a,b), \mathbb{P}[X=a, Y=b]) : a \in \mathscr{A},\ b \in \mathscr{B}\}$, where $\mathscr{A}$ is the set of all possible values taken by X and $\mathscr{B}$ is the set of all possible values taken by Y.*

Given a joint distribution of X and Y, the distribution $\mathbb{P}[X=a]$ of X is called the *marginal distribution* of X, and can be found by summing over the values of Y. That is,

$$\mathbb{P}[X=a] = \sum_{b \in \mathscr{B}} \mathbb{P}[X=a, Y=b].$$

The marginal distribution of Y is defined analogously.

A joint distribution over any set of random variables $X_1, \ldots, X_n$ (for example, X_i could be the value of the ith roll of a sequence of n die rolls) is $\mathbb{P}[X_1 = a_1, \ldots, X_n = a_n]$, where $a_i \in \mathscr{A}_i$ and $\mathscr{A}_i$ is the set of possible values for X_i. The marginal distribution of X_i can be obtained by summing over all the possible values of the other variables.

Independence

Independence for random variables is defined in an analogous fashion to independence for events:

Definition 16.4 (Independence)**.** *Random variables X and Y on the same probability space are said to be* independent *if the events $X = a$ and $Y = b$ are independent for all values a, b. Equivalently, the joint distribution of independent r.v.'s decomposes as*

$$\mathbb{P}[X = a, Y = b] = \mathbb{P}[X = a]\,\mathbb{P}[Y = b], \quad \forall a, b.$$

Mutual independence of more than two r.v.'s is defined similarly.

A very important example of independent random variables are indicator random variables for independent events. If I_i denotes the indicator r.v. for the ith toss of a coin being H, then $I_1, \ldots, I_n$ are mutually independent random variables. This example motivates the commonly used phrase "*independent and identically distributed (i.i.d.)* set of random variables." In this example, $\{I_1, \ldots, I_n\}$ is a set of i.i.d. indicator random variables.

Exchangeability

Suppose we randomly sample *without replacement* $n \leq N$ balls from an urn containing B black balls and $N - B$ white balls. We showed earlier that the number Y of black balls in the sample follows the Hypergeometric(N, B, n) distribution. Note that we can write

$$Y = I_1 + \cdots + I_n,$$

where I_i is the indicator random variable that equals 1 if the ith draw is black and 0 otherwise.

By the argument used to derive (2), we have (here $\sum_{i=1}^n a_i$ plays the role of k in (2)):

$$\mathbb{P}[I_1 = a_1, \ldots, I_n = a_n] = \frac{\binom{B}{\sum_{i=1}^n a_i}\binom{N-B}{n-\sum_{i=1}^n a_i}}{\binom{N}{n}} \cdot \frac{1}{\binom{n}{\sum_{i=1}^n a_i}}. \tag{4}$$

An important feature of (4) is that the right-hand side depends on $(a_1, \ldots, a_n)$ only through their sum $\sum_{i=1}^n a_i$. In particular, the joint distribution is invariant under permutations of the indices. That is, for any permutation π of $\{1, \ldots, n\}$,

$$\mathbb{P}[I_{\pi(1)} = a_1, \ldots, I_{\pi(n)} = a_n] = \mathbb{P}[I_1 = a_1, \ldots, I_n = a_n] \tag{5}$$

for all $a_1, \ldots, a_n \in \{0, 1\}$.

Random variables satisfying (5) are called *exchangeable*.

Definition 16.5 (Exchangeability)**.** *A collection of random variables $X_1, \ldots, X_n$ with common range $\mathscr{A}$ is said to be* exchangeable *if $(X_{\pi(1)}, \ldots, X_{\pi(n)})$ has the same joint distribution as $(X_1, \ldots, X_n)$ for every permutation π of $\{1, \ldots, n\}$.*

Thus, $I_1, \ldots, I_n$ are exchangeable random variables.

We now record several consequences of exchangeability.

1. **Identical marginals.** The random variables $X_1, \ldots, X_n$ have identical marginal distributions:

$$\mathbb{P}[X_1 = a] = \mathbb{P}[X_2 = a] = \cdots = \mathbb{P}[X_n = a] \tag{6}$$

   for all $a \in \mathscr{A}$.

   This follows because, for any i and any $a \in \mathscr{A}$,

$$\begin{aligned}\mathbb{P}[X_i = a] &= \sum_{b_1,\ldots,b_{i-1},b_{i+1},\ldots,b_n \in \mathscr{A}} \mathbb{P}[X_1 = b_1, \ldots, X_{i-1} = b_{i-1}, X_i = a, X_{i+1} = b_{i+1}, \ldots, X_n = b_n] \\ &= \sum_{b_1,\ldots,b_{i-1},b_{i+1},\ldots,b_n \in \mathscr{A}} \mathbb{P}[X_1 = a, \ldots, X_{i-1} = b_{i-1}, X_i = b_1, X_{i+1} = b_{i+1}, \ldots, X_n = b_n] \\ &= \mathbb{P}[X_1 = a],\end{aligned}$$

   where we used exchangeability in the second step.

   Although the X_i have identical distributions, they are generally *not independent*. In fact, i.i.d. implies exchangeability, but the converse need not hold.

2. **Identical pairwise distributions.** For any distinct i, j,

$$\mathbb{P}[X_i = a, X_j = b] = \mathbb{P}[X_1 = a, X_2 = b] \tag{7}$$

   for all $a, b \in \mathscr{A}$. The proof follows by summing over the remaining variables and applying exchangeability.

3. **Identical k-tuple distributions.** More generally, for any distinct indices $i_1, \ldots, i_k$, where $k < n$, the joint distribution of $(X_{i_1}, \ldots, X_{i_k})$ is the same as that of $(X_1, \ldots, X_k)$. The proof is similar to the previous cases.

4 Expectation

The distribution of a r.v. contains *all* the information about the r.v. In most applications, however, the complete distribution of a r.v. is very hard to calculate. For example, consider the homework permutation example with $n = 20$. In principle, we would 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 (though in practice we could streamline this calculation a bit)! 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 *summarize* 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 16.6 (Expectation)**.** *The expectation of a discrete random variable X is defined as*

$$\mathbb{E}[X] = \sum_{a \in \mathscr{A}} a \times \mathbb{P}[X = a], \tag{8}$$

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

Technical Note. Expectation is well defined provided that the sum on the right hand side of (8) is absolutely convergent, i.e., $\sum_{a \in \mathscr{A}} |a| \times \mathbb{P}[X = a] < \infty$. There are discrete random variables for which expectations do not exist, such as the r.v. X with distribution $\mathbb{P}[X = i] = \frac{6}{\pi^2 i^2}$ for all positive integers i. (The reason for the factor $\frac{6}{\pi^2}$ here is to make the probabilities sum to 1, since $\sum_{i=1}^{\infty} \frac{1}{i^2} = \frac{\pi^2}{6}$.)

For our simpler permutation example with only 3 students, the expectation is

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

That is, 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 $\mathbb{E}[X]$ may not actually be a value that X can take). 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 is a physical interpretation of the expectation of a random variable: imagine carving out a wooden cutout figure of the probability distribution as in Figure 4. Then the expected value of the distribution is the balance point (directly below the center of gravity) of this object.

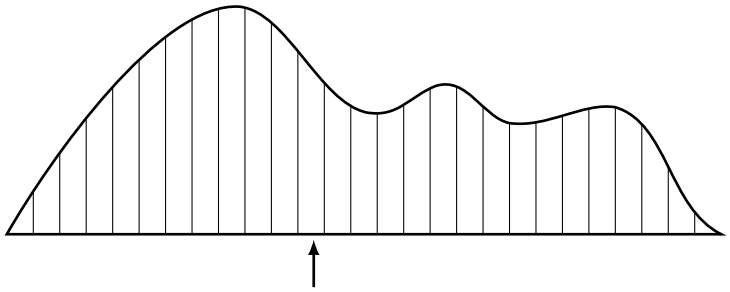

Figure 4: Physical interpretation of expected value as the balance point.

4.1 Examples

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

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

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

2. **Two dice.** Throw two fair dice and 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
$\mathbb{P}[X = a]$	$\frac{1}{36}$	$\frac{1}{18}$	$\frac{1}{12}$	$\frac{1}{9}$	$\frac{5}{36}$	$\frac{1}{6}$	$\frac{5}{36}$	$\frac{1}{9}$	$\frac{1}{12}$	$\frac{1}{18}$	$\frac{1}{36}$

   The expectation is therefore

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

3. **Roulette.** A roulette wheel is spun (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). 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 $\mathbb{P}[X = 1] = \frac{18}{38}$, $\mathbb{P}[X = -1] = \frac{20}{38}$. Thus,
$$\mathbb{E}[X] = \left(1 \times \frac{18}{38}\right) + \left(-1 \times \frac{20}{38}\right) = -\frac{1}{19};$$
i.e., you expect to lose about a nickel per game. Notice how the zeros tip the balance in favor of the casino!

4.2 Linearity of Expectation

So far, we have 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 16.1. *For any two random variables X and Y on the same probability space, we have*
$$\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y].$$
Also, for any constant c, we have
$$\mathbb{E}[cX] = c\,\mathbb{E}[X].$$

Proof. We first rewrite the definition of expectation in a more convenient form. Recall from Definition 16.6 that
$$\mathbb{E}[X] = \sum_{a \in \mathscr{A}} a \times \mathbb{P}[X = a].$$
Consider a particular term $a \times \mathbb{P}[X = a]$ in the above sum. Notice that $\mathbb{P}[X = a]$, by definition, is the sum of $\mathbb{P}[\omega]$ over those sample points ω for which $X(\omega) = a$. Furthermore, 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
$$\mathbb{E}[X] = \sum_{\omega \in \Omega} X(\omega) \times \mathbb{P}[\omega]. \tag{9}$$
This equivalent definition of expectation will make the present proof much easier (though it is usually less convenient for actual calculations). Applying (9) to $\mathbb{E}[X+Y]$ gives:
$$\begin{aligned}
\mathbb{E}[X+Y] &= \sum_{\omega \in \Omega} (X+Y)(\omega) \times \mathbb{P}[\omega] \\
&= \sum_{\omega \in \Omega} (X(\omega) + Y(\omega)) \times \mathbb{P}[\omega] \\
&= \sum_{\omega \in \Omega} (X(\omega) \times \mathbb{P}[\omega]) + \sum_{\omega \in \Omega} (Y(\omega) \times \mathbb{P}[\omega]) \\
&= \mathbb{E}[X] + \mathbb{E}[Y]
\end{aligned}$$
In the last step, we used (9) 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 16.1 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. We can use Theorem 16.1 to conclude things like $\mathbb{E}[3X-5Y] = 3\mathbb{E}[X]-5\mathbb{E}[Y]$, regardless of whether or not X and Y are independent. This important property is known as linearity of expectation.

Important caveat: Theorem 16.1 does ***not*** *say that* $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$*, or that* $\mathbb{E}\left[\frac{1}{X}\right] = \frac{1}{\mathbb{E}[X]}$*, etc. These claims are not true in general. It is only* sums *and* differences *and* constant multiples *of random variables that behave so nicely.*

4.3 Applications of Linearity of Expectation

Now let us see some examples of Theorem 16.1 in action.

1. **Two dice again.** Here is a much less painful way of computing $\mathbb{E}[X]$, where X is the sum of the scores of the two dice. Note that $X = Y_1 + Y_2$, where Y_i is the score on die i. We know from example 1 in Section 4.1 that $\mathbb{E}[Y_1] = \mathbb{E}[Y_2] = \frac{7}{2}$. So, by Theorem 16.1, we have $\mathbb{E}[X] = \mathbb{E}[Y_1] + \mathbb{E}[Y_2] = 7$.

2. **More roulette.** Suppose we play the roulette game mentioned in Section 4.1 $n \geq 1$ times. Let X_n be our expected net winnings. Then $X_n = Y_1 + Y_2 + \cdots + Y_n$, where Y_i is our net winnings in the ith play. We know from earlier that $\mathbb{E}[Y_i] = -\frac{1}{19}$ for each i. Therefore, by Theorem 16.1, $\mathbb{E}[X_n] = \mathbb{E}[Y_1] + \mathbb{E}[Y_2] + \cdots + \mathbb{E}[Y_n] = -\frac{n}{19}$. For $n = 1000$, $\mathbb{E}[X_n] = -\frac{1000}{19} \approx -53$, so if you play 1000 games, you expect to lose about \$53.

3. **Fixed points of permutations.** Let us return to the homework permutation example with an arbitrary number n of students. Let X_n denote the number of students who receive their own homework after shuffling (or equivalently, the number of fixed points). To take advantage of Theorem 16.1, we need to write X_n as a *sum* of simpler r.v.'s. Since X_n *counts* the number of times something happens, we can write it as a sum using the following useful trick:

$$X_n = I_1 + I_2 + \cdots + I_n, \qquad \text{where } I_i = \begin{cases} 1, & \text{if student } i \text{ gets their own homework,} \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

   [You should think about this equation for a moment. Remember that all the I_i's are random variables. What does an equation involving random variables mean? What we mean is that, *at every sample point* ω, we have $X_n(\omega) = I_1(\omega) + I_2(\omega) + \cdots + I_n(\omega)$. Why is this true?]

   A Bernoulli random variable such as I_i is called an indicator random variable of the corresponding event (in this case, the event that student i gets their own homework). For indicator r.v.'s, the expectation is particularly easy to calculate. Specifically,

$$\mathbb{E}[I_i] = (0 \times \mathbb{P}[I_i = 0]) + (1 \times \mathbb{P}[I_i = 1]) = \mathbb{P}[I_i = 1].$$

   In our case, we have

$$\mathbb{P}[I_i = 1] = \mathbb{P}[\text{student } i \text{ gets their own homework}] = \frac{1}{n}.$$

   We can now apply Theorem 16.1 to (10), yielding

$$\mathbb{E}[X_n] = \mathbb{E}[I_1] + \mathbb{E}[I_2] + \cdots + \mathbb{E}[I_n] = n \times \frac{1}{n} = 1.$$

   So, we see that the expected number of students who get their own homeworks in a class of size n is 1. That is, the expected number of fixed points in a random permutation of n items is always 1, regardless of n!

4. **Coin tosses.** Toss a fair coin $n \geq 1$ times. Let the r.v. X_n be the number of heads observed. As in the previous example, to take advantage of Theorem 16.1 we write

$$X_n = I_1 + I_2 + \cdots + I_n,$$

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

$$\mathbb{E}[I_i] = \mathbb{P}[I_i = 1] = \mathbb{P}[i\text{th toss is } H] = \frac{1}{2}.$$

Using Theorem 16.1, we therefore get

$$\mathbb{E}[X_n] = \sum_{i=1}^{n} \frac{1}{2} = \frac{n}{2}.$$

More generally, in n tosses of a biased coin that comes up H with probability p, $\mathbb{E}[X_n] = np$. (Check this!) So the expectation of a binomial r.v. $X \sim \text{Bin}(n,p)$ is equal to np. Note that it would have been much messier (though possible) to reach the same conclusion by computing this directly from the definition of expectation in (8) and the distribution of a binomial r.v. in (1).

5. **Sampling without replacement.** Consider sampling *without replacement* $n \leq N$ balls from an urn containing B black balls and $N - B$ white balls. Let Y denote the number of black balls in the sample. We know that Y follows the Hypergeometric(N,B,n) distribution. However, computing $\mathbb{E}[Y]$ directly from the distribution (see (3)) can be cumbersome. A much simpler approach is to use indicator random variables. Write

$$Y = I_1 + \cdots + I_n,$$

where I_i is the indicator that the ith draw is black. Although the random variables $I_1, \ldots, I_n$ are *not* independent, linearity of expectation still applies:

$$\mathbb{E}[Y] = \mathbb{E}[I_1] + \cdots + \mathbb{E}[I_n].$$

Moreover, by exchangeability, the I_i have identical distributions, so

$$\mathbb{E}[Y] = n\,\mathbb{E}[I_1].$$

Finally, since the probability that the first draw is black is $\mathbb{P}[I_1 = 1] = \frac{B}{N}$, we conclude that

$$\mathbb{E}[Y] = n\frac{B}{N}.$$

6. **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 m tosses of a biased coin with $\mathbb{P}[H] = \frac{1}{n}$ (why?). So, from the previous example, we get $\mathbb{E}[X] = \frac{m}{n}$. 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 would get into a messy calculation involving binomial coefficients.

   Here is another example on the same sample space. Let the r.v. Y_n be the number of empty bins. The distribution of Y_n is horrible to contemplate: to get a feel for this, you might like to write it down for $m = n = 3$ (i.e., 3 balls, 3 bins). However, computing the expectation $\mathbb{E}[Y_n]$ is easy using Theorem 16.1. As in the previous two examples, we write

$$Y_n = I_1 + I_2 + \cdots + I_n, \tag{11}$$

where I_i is the indicator r.v. of the event "bin i is empty". The expectation of I_i is easy to find:

$$\mathbb{E}[I_i] = \mathbb{P}[I_i = 1] = \mathbb{P}[\text{bin } i \text{ is empty}] = \left(1 - \frac{1}{n}\right)^m,$$

as discussed earlier. Applying Theorem 16.1 to (11), we therefore obtain

$$\mathbb{E}[Y_n] = \sum_{i=1}^{n} \mathbb{E}[I_i] = n\left(1 - \frac{1}{n}\right)^m,$$

a simple formula, quite easily derived. Let us see how it behaves in the special case $m = n$ (same number of balls as bins). In this case we get $\mathbb{E}[Y_n] = n\left(1 - \frac{1}{n}\right)^n$. Now the quantity $\left(1 - \frac{1}{n}\right)^n$ can be approximated (for large enough values of n) by the number $\frac{1}{e}$.[2] So we see that, for large n,

$$\mathbb{E}[Y_n] \approx \frac{n}{e} \approx 0.368n.$$

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

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

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

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 $(1 - \frac{1}{n})^n \approx 0.358$, which is very close to $\frac{1}{e} \approx 0.368$. The approximation gets better and better for larger n.