Queen's University
School of Computing

CISC 203: Discrete Mathematics for Computing II
Module 9: Probability (Part II)
Fall 2021

This module corresponds to the following sections from your textbook:

32. Conditional Probability and Independence
33. Random Variables
34. Expectation

1 Conditional Probability

Let us introduce the concept of conditional probability with the following xkcd comic set in the middle of an electrical storm. The hiker on the left expresses concern and wishes to go inside. But the hiker on the right dismisses the concern, explaining that the chances of an American dying in a thunderstorm is about one in seven million.

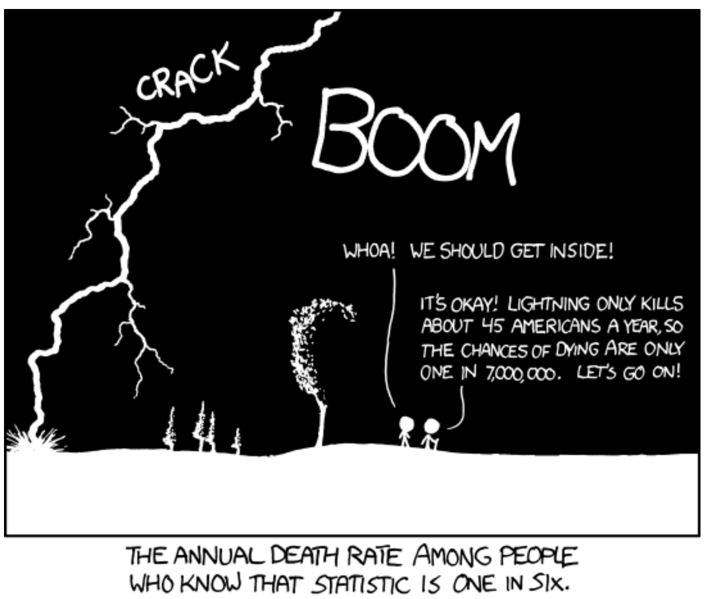


Figure 1: Source: https://xkcd.com/795/. The caption is reminiscent of a famous quote from Alexander Pope: "A little knowledge is a dangerous thing".

Let A represent the event that an American dies by lightning. Let B represent the event that an American deliberately walks into an electrical storm. The probability of each of these events occurring, $P(A)$ and $P(B)$, respectively, are very low. However, the hiker on the right is ignorant of conditional probability, and fails to distinguish between:

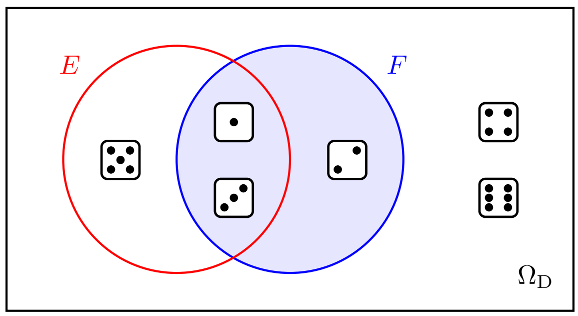
1. The probability of an American dying by lightning, i.e., $P(A)$; and

2. The **conditional probability** of an American dying by lightning **given that they deliberately walk into an electrical storm**, denoted $P(A \mid B)$ (i.e., the probability that the event A occurs given that the event B occurs).

We can see intuitively that in the example above the conditional probability $P(A \mid B)$ should be much higher than the probability $P(A)$.

Example 1. Recall the sample space for rolling a die, $S_{\mathrm{D}} = \{$⚀, ⚁, ⚂, ⚃, ⚄, ⚅$\}$. Let E denote the event "rolled an odd number", and F denote the event "rolled a number less than 4". A friend rolls the die and tells you that they saw a number less than 4 (i.e., F occurred). What is the probability that they also rolled an odd number (i.e., the probability of E occurring, given that F occurred)?

Let's look at this scenario diagrammatically, where E and F are represented by circles and each possible outcome is placed in the appropriate location.



Immediately, we see that rolling a 4 or a 6 means neither of the events E nor F occur, so both of those outcomes lie outside of each circle. We also see that both circles contain three outcomes, so we have

$$P(E) = \frac{3}{6} = \frac{1}{2} \quad \text{and} \quad P(F) = \frac{3}{6} = \frac{1}{2}.$$

However, we don't care about all possible outcomes, since we already know that F occurred. That is, we already know our friend rolled a number less than 4, which narrows the sample space down to $\{$⚀, ⚁, ⚂$\}$. Of these outcomes, the only rolls that would make E occur are ⚀ and ⚂. Therefore, we conclude that the probability of E occurring given that F occurred, denoted $P(E \mid F)$, is $2/3$.

To summarize how we calculated $P(E \mid F)$ in the above die-rolling example: we considered the number of outcomes shared by both events E and F, and divided that by the number of outcomes in the event F that is known to have occurred. Note that **this approach assumes the use of a fair die where each outcome is equally probable**. More generally, we calculate conditional probabilities using the following formula.

Definition 2 (Conditional probability)**.** Let A and B be events in a sample space (S, P) and suppose $P(B) \neq 0$. The conditional probability $P(A \mid B)$, the probability of A given B, is

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

Example 3. Let b be a randomly-generated bit string of length four, where each bit in the sequence is equally likely to be generated as a `0` or `1`. Let E be the event "b contains two consecutive `1`s" and F be the event "b both starts and ends with a `1`".

We have $P(E) = 1/2$, since the bit strings containing consecutive `1`s are

$$\{\texttt{0011}, \texttt{0110}, \texttt{1100}, \texttt{0111}, \texttt{1110}, \texttt{1011}, \texttt{1101}, \texttt{1111}\}.$$

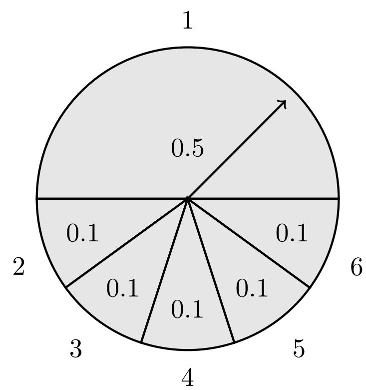
Since the first and last bits of b each have a $1/2$ probability of being generated as a `1`, by the multiplication principle we have $P(F) = 1/4$. The set of bit strings that start and end with a `1` is

$$\{\texttt{1001}, \texttt{1011}, \texttt{1101}, \texttt{1111}\}.$$

Assume that F occurred. What is $P(E \mid F)$? We know that $P(E \cap F) = 3/16$, since the set of bit strings common to both events E and F is $\{\texttt{1011}, \texttt{1101}, \texttt{1111}\}$. We also know that $P(F) = 1/4$. So, using the conditional probability formula, we get

$$P(E \mid F) = \frac{3/16}{1/4} = \frac{3}{4}.$$

Example 4. Consider the following spinner, where the arrow represents a needle that can be spun around. The set of all possible outcomes is $S = \{1, 2, 3, 4, 5, 6\}$ and the probability function $P : S \longrightarrow \mathbb{R}$ is defined by $P(1) = 0.5$ and $P(2) = P(3) = P(4) = P(5) = P(6) = 0.1$.



Let $A = \{1, 2, 3\}$. We can immediately see that $P(A) = 1/2 + 1/10 + 1/10 = 7/10$.

Now, suppose that we have spun the needle.

Suppose that the event $B = \{2, 3\}$ occurred. Both of the outcomes in B are also in A, so in this case we know that A also occurred. So, we have $P(A \mid B) = 1$. We can also calculate this as follows:

$$\begin{aligned} P(B) &= \frac{2}{10} \\ P(A \cap B) &= \frac{2}{10} \\ P(A \mid B) &= \frac{2/10}{2/10} = 1 \end{aligned}$$

Suppose that the event $C = \{4, 6\}$ occurred. Neither of the outcomes in C are in A, so in this case we know that A did not occur. So, we have $P(A \mid C) = 0$. We can also calculate this as follows:

$$\begin{aligned} P(C) &= \frac{2}{10} \\ P(A \cap C) &= 0 \\ P(A \mid C) &= \frac{0}{2/10} = 0 \end{aligned}$$

Suppose that the event $D = \{1, 2, 4\}$ occurred. Of the three outcomes in D, 1 and 2 are also in A, so in this case we know that A might have occurred. But it would be wrong to say that $P(A \mid D) = 2/3$, since outcome 1 has a much higher probability of occurring than 2 or 4. Instead, we must calculate $P(A \mid D)$ as follows:

$$\begin{aligned} P(D) &= \frac{1}{2} + \frac{1}{10} + \frac{1}{10} = \frac{7}{10} \\ P(A \cap D) &= \frac{1}{2} + \frac{1}{10} = \frac{6}{10} \\ P(A \mid D) &= \frac{6/10}{7/10} = \frac{6}{7} \end{aligned}$$

Exercise

Is it possible to have events A and B where $P(A) = 0.3$, $P(B) = 0.1$, and $P(B \mid A) = 0.5$? Hint: What is the largest possible value we could have for $P(A \cap B)$?

1.1 Independence

A classic illustration of independence is given by the "gambler's fallacy". Imagine a gambler in a casino who is rolling a fair die, and has rolled a ⚅ three times in a row. The gambler might think that it's unlikely that the next roll of the die will produce a ⚅ again, because across many rolls the fair die should produce a roughly equal number of each outcome. This is a fallacy, because the die has no "memory" of previous rolls. Each time the gambler rolls the die, each of the six possible outcomes have a $1/6$ chance of occurring, regardless of the outcome of the previous rolls.

Example 5. A coin is flipped five times. What is the probability that the first flip shows T given that exactly three H are flipped?

Let A be the event that the first flip produces T and B be the event that exactly three of the flips produce H. The set $A \cap B$ contains the $\binom{4}{3}$ sequences where the first flip produces T and exactly three of the four remaining flips produce H. So, we have

$$P(A) = \frac{2^4}{2^5} = \frac{1}{2}, \quad P(B) = \frac{\binom{5}{3}}{2^5} = \frac{10}{32} = \frac{5}{16}, \quad \text{and} \quad P(A \cap B) = \frac{\binom{4}{3}}{2^5} = \frac{4}{32} = \frac{1}{8}.$$

So,

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{1/8}{5/16} = \frac{2}{5}.$$

Observe that $P(A \mid B)$ is lower than $P(A)$. That is because if the experiment is performed and we learn that B occurred (i.e., three H were produced), this reduces the probability that A occurred (i.e., the first flip producing a T). Since knowledge about B tells us something about A (and vice-versa), we say that A and B are **dependent**.

Now, let C be the event that the first flip produces H and D be the event that the last flip produces H. So, we have

$$P(C) = P(D) = \frac{2^4}{2^5} = \frac{1}{2} \quad \text{and} \quad P(C \cap D) = \frac{2^3}{2^5} = \frac{1}{4}.$$

Therefore,

$$P(C \mid D) = \frac{P(C \cap D)}{P(D)} = \frac{1/4}{1/2} = \frac{1}{2} = P(C).$$

C and D tell us something about the first flip and the last flip, respectively. But learning the outcome of the first flip tells us absolutely nothing about the outcome of the last flip, because the coin has no "memory" as explained prior to this example. Therefore, we have $P(C \mid D) = P(C)$ and we say that C and D are **independent**.

Definition 6 (Independent events)**.** Let A and B be events in a sample space. We say that these events are independent, provided

$$P(A \cap B) = P(A)P(B).$$

Equivalently, $P(A \mid B) = P(A)$ and $P(B \mid A) = P(B)$ for independent events A and B.

Example 7. A bag containing twenty marbles: ten are coloured red and ten are coloured blue.

Consider the experiment where we (i) draw one marble from the bag and check its colour, (ii) put the marble back in the bag, and (iii) draw a second marble and check its colour. Let A be the event that the first marble is red, and B be the event that the second marble is red. Are A and B independent?

There are $20 \times 20 = 400$ outcomes to the experiment, and each outcome is equally likely. There are $10 \times 20 = 200$ outcomes where the first marble drawn is red. So, $P(A) = \frac{200}{400} = \frac{1}{2}$. Similarly, $P(B) = \frac{1}{2}$. There are $10 \times 10 = 100$ outcomes where both the first marble is red and the second marble is red, so $P(A \cap B) = \frac{100}{400} = \frac{1}{4}$. So, we have

$$P(A \cap B) = \frac{1}{4} = \frac{1}{2} \times \frac{1}{2} = P(A)P(B).$$

So, we conclude that A and B are independent events.

Exercise

Let us remove the second step (ii) of the experiment, so that the first marble is **not** placed back into the bag. Now, are A and B independent? Show why or why not, by finding $P(A)$, $P(B)$, and $P(A \cap B)$ as was done above. Cross-check your answer on p. 225–226 of the textbook.

2 Random Variables

Thus far, when discussing events from a sample space, we had to explicitly denote each event. For example, when rolling two dice, we may want to define the event A as "the sum of the rolled dice equals 2", B as "the sum of the rolled dice equals 3", and so on, up to the event K as "the sum of the rolled dice equals 12". This can get tedious if we have many events to define. The following example builds up to a new concept that will make it much more convenient to refer to these events A through K without having to individually denote each of them.

Example 8. As we have seen before, let (S_{2D}, P_{2D}) be the sample space for the experiment of rolling two dice, with

$$S_{\text{2D}} = \{\text{⚀⚀}, \text{⚀⚁}, \text{⚀⚂}, \ldots, \text{⚅⚄}, \text{⚅⚅}\}$$

and $P_{2D}(s) = 1/36$ for all $s \in S_{2D}$.

Let $X : S \longrightarrow \mathbb{N}$ be a function, defined on the sample space (S_{2D}, P_{2D}), that maps each outcome in $s \in S_{2D}$ to a value $x \in \mathbb{N}$ where x is the sum of the values on both dice. To illustrate, we list $X(s)$ for all possible $s \in S_{2D}$ as follows:

$X(\text{⚀⚀}) = 2$					
$X(\text{⚀⚁}) = 3$	$X(\text{⚁⚀}) = 3$				
$X(\text{⚀⚂}) = 4$	$X(\text{⚁⚁}) = 4$	$X(\text{⚂⚀}) = 4$			
$X(\text{⚀⚃}) = 5$	$X(\text{⚁⚂}) = 5$	$X(\text{⚂⚁}) = 5$	$X(\text{⚃⚀}) = 5$		
$X(\text{⚀⚄}) = 6$	$X(\text{⚁⚃}) = 6$	$X(\text{⚂⚂}) = 6$	$X(\text{⚃⚁}) = 6$	$X(\text{⚄⚀}) = 6$	
$X(\text{⚀⚅}) = 7$	$X(\text{⚁⚄}) = 7$	$X(\text{⚂⚃}) = 7$	$X(\text{⚃⚂}) = 7$	$X(\text{⚄⚁}) = 7$	$X(\text{⚅⚀}) = 7$
$X(\text{⚁⚅}) = 8$	$X(\text{⚂⚄}) = 8$	$X(\text{⚃⚃}) = 8$	$X(\text{⚄⚂}) = 8$	$X(\text{⚅⚁}) = 8$	
$X(\text{⚂⚅}) = 9$	$X(\text{⚃⚄}) = 9$	$X(\text{⚄⚃}) = 9$	$X(\text{⚅⚂}) = 9$		
$X(\text{⚃⚅}) = 10$	$X(\text{⚄⚄}) = 10$	$X(\text{⚅⚃}) = 10$			
$X(\text{⚄⚅}) = 11$	$X(\text{⚅⚄}) = 11$				
$X(\text{⚅⚅}) = 12$					

To be more formal, let us redefine S_{2D} as a set of ordered pairs of numbers instead; i.e., we will represent ⚀⚀ as the list $(1, 1)$, ⚅⚃ as the list $(6, 4)$, and so on.

Now, we can use the function X as a "variable" by referring to events as instances of "X takes on a particular value $x \in \mathbb{N}$". Recalling the events A through K we defined immediately prior to this example, the event $X = 2$ is equivalent to A, the event $X = 3$ is equivalent to B, and the event $X = 12$ is equivalent to K.

We call this special function X a **random variable**. But despite the name, do not lose sight of the fact that X is actually a function. So for example, when we refer to an event $X = 4$, we are referring to the set containing the outcomes $s_i \in S_{2D}$ where $X(s_i) = 4$.

In other words, $X = 4$ is shorthand for the event (i.e., set of outcomes)

$$\{x \in S : X(s) = 4\},$$

which in this case is the event

{⚀⚂, ⚁⚁, ⚂⚀},

or, using our more formal notation,

$$\{(1, 3), (2, 2), (3, 1)\}.$$

We may also refer to the event $X > 10$, which is the event defined by the set of outcomes where the sum of the dice is greater than 10,

$$\{(5, 6), (6, 5), (6, 6)\}.$$

We define random variables more formally as follows.

Definition 9 (Random variable)**.** A random variable is a function defined on a probability space; that is, if (S, P) is a sample space, then a random variable is a function $X : S \longrightarrow V$ for some set V.

In Definition 9, the set V is typically a set of numbers, such as $\mathbb{N}$ (as in Example 8 above), $\mathbb{Z}$, or $\mathbb{R}$. However, V can be a non-numeric set as well.

Example 10. Let (S, P) be the sample space representing the experiment where a card is drawn at random from a deck, and $X : S \longrightarrow V$ be the random variable that maps each outcome to the suit $v \in V$ of the card where $V = \{\clubsuit, \heartsuit, \diamondsuit, \spadesuit\}$. For example, $X(10\clubsuit) = \clubsuit$, $X(2\clubsuit) = \clubsuit$ and $X(9\spadesuit) = \spadesuit$.

Example 11. Let (S, P) be the sample space representing ten tosses of a fair coin, and $X : S \longrightarrow \mathbb{Z}$ be the random variable that maps each outcome to the number of H minus the number of T. For example,

$$X(\text{HHHHTTTTTT}) = -2.$$

We may also define $X_H : S \longrightarrow \mathbb{N}$ as the random variable that maps each outcome to the number of H, e.g.,

$$X_H(\text{HHHHTTTTTT}) = 4.$$

And we may define $X_T : S \longrightarrow \mathbb{N}$ as the random variable that maps each outcome to the number of T, e.g.,

$$X_T(\text{HHHHTTTTTT}) = 6.$$

Note that we have $X = X_H - X_T$, which means that for any $s \in S$, we have $X(s) = X_H(s) - X_T(s)$.

Now let us calculate the probability of flipping at least four H and at least four T. We may express this as $P(X_H \geq 4 \cap X_T \geq 4)$. Or equivalently, $P(4 \leq X_H \leq 6)$. So, we have

$$\begin{aligned} P(4 \leq X_H \leq 6) &= P(X = 4) + P(X = 5) + P(X = 6) \\ &= \frac{\binom{10}{4}}{2^{10}} + \frac{\binom{10}{5}}{2^{10}} + \frac{\binom{10}{6}}{2^{10}} = \frac{672}{1024} = \frac{21}{32}. \end{aligned}$$

Example 12 (Binomial random variable)**.** Suppose an experiment with a rigged coin that produces H with probability p and T with probability $1-p$. The coin is flipped n times. Let X denote the number of times that we see H.

For an integer $h \in \mathbb{Z}$, what is $P(X = h)$? The event $X = h$ is impossible for $h < 0$ or $h > n$, so in those cases we have $P(X = h) = 0$.

For $0 \leq h \leq n$, there are exactly $\binom{n}{h}$ outcomes of n flips where exactly h coins show H. All of these outcomes have the probability $p^h(1-p)^{n-h}$ of occurring. Therefore, we have

$$P(X = h) = \binom{n}{h} p^h (1-p)^{n-h}.$$

X is called a **binomial random variable**, since the expression for $P(X = h)$ can be derived by expanding $(p+q)^n$ using the binomial theorem and substituting $q = 1-p$ into one of the resulting terms in the expansion, $\binom{n}{h} p^h q^{n-h}$.

Exercise

Suppose we are tossing a coin that produces H with probability $1/3$ and T with probability $2/3$. Let (S, P) be the sample space for the experiment of tossing the coin four consecutive times and $X(s)$ be the random variable that maps each outcome $s \in S$ to the number of H that is produced in the outcome s. What is $P(X = 2)$?

2.1 Independent Random Variables

Recall that we say two events A and B defined on a sample space (S, P) are independent if $P(A \cap B) = P(A)P(B)$. This means that knowing A occurred does not tell us anything about whether B occurred, and vice-versa; i.e., $P(B \mid A) = P(B)$ and $P(A \mid B) = P(A)$. We can analogously define independence for random variables in the same way, as follows.

Definition 13 (Independent random variables)**.** We say that the random variables X_1 and X_2 are independent if, for all possible values for a and b, we have

$$P(X_1 = a \text{ and } X_2 = b) = P(X_1 = a)P(X_2 = b).$$

Example 14. Recall the pair-of-dice sample space (S_{2D}, P_{2D}). We define the random variable $X_1(s)$ to give the value of the first die in an outcome $s \in S_{2D}$, and the random variable $X_2(s)$ to give the value of the second die in an outcome $s \in S_{2D}$. For example, $X_1($⚄⚂$) = 5$ and $X_2($⚄⚂$) = 3$. Finally, let $X = X_1 + X_2$, e.g., $X($⚄⚂$) = 8$.

Since all outcomes are equally likely, we know that $P(X = 5) = 4/36 = 1/9$.

Suppose that we perform the experiment and obtain the outcome s. Intuitively, we can see that knowledge of $X_2(s)$ gives us some knowledge about X. For example, if $X_2(s) = 4$, this tells us that $X(s) \neq 4$. In fact, we can say that $P(X = 4 \mid X_2 = 4) = 0$. We can also say that $P(X = 5 \mid X_2 = 4) = 1/6$, since only one out of the six outcomes for the other die would result in the event $X = 5$. More formally,

$$P(X = 5 \mid X_2 = 4) = \frac{P(X = 5 \text{ and } X_2 = 4)}{P(X_2 = 4)} = \frac{1/36}{1/6} = \frac{1}{6} \neq P(X = 5).$$

We can also see intuitively that knowledge of $X_2(s)$ tells us nothing about $X_1(s)$, since the value of one die is not influenced by the other. This is confirmed by

$$P(X_1 = a \mid X_2 = b) = \frac{P(X_1 = a \text{ and } X_2 = b)}{P(X_2 = b)} = \frac{1/36}{1/6} = \frac{1}{6} = P(X_1 = a),$$

where a and b are integers from 1 to 6.

3 Expectation and Variance

We informally introduce the idea of expectation and variance as follows. Consider an experiment with a sample space (S, P) and a random variable X that maps each outcome $s \in S$ to a numeric value $v \in V$. If we repeat the experiment many times and write down the value $X(s_i)$ where s_i is the outcome of each experiment, what would the average value of X be (i.e., what is our **expectation** for the value of X)? How widely spread might the values of X be from the average value (i.e., what is the **variance** of X)? In this section, we will learn methods of measuring both of these values.

Note that in some cases such as Example 10, the random variable X gives non-numeric values, and so the concept of an "average" value would not be meaningful in such cases.

3.1 Expectation

The "average" of X is more commonly known as the **expectation** or **expected value**, and it is denoted by $\mathrm{E}(X)$. The expectation of a random variable is the weighted average of the values produced by the random variable on each outcome, where each weight is determined by the probability of that outcome.

Definition 15 (Expectation of a random variable)**.** Let X be a real-valued random variable defined on a sample space (S, P). The **expectation** (or the **expected value**) of X is

$$\mathrm{E}(X) = \sum_{s \in S} X(s)P(s).$$

Example 16. Assume we have two dice: die d_1 is a fair six-sided die, and die d_2 is a biased six-sided die where P[{⚀}] = 0.75 and $P(\{$⚁$\}) = P(\{$⚂$\}) = P(\{$⚃$\}) = P(\{$⚄$\}) = P(\{$⚅$\}) = 0.05$. Let X_1 and X_2 denote the random variables associated with rolling die d_1 and die d_2, respectively, which map each outcome to the numeric value of the rolled die. The average value we would expect to roll with die d_1 is

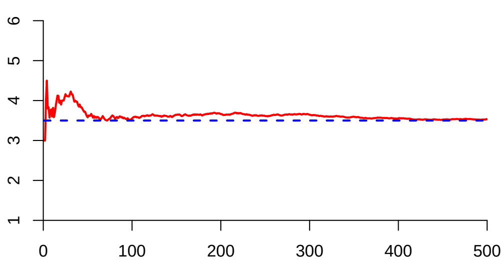
$$\mathrm{E}(X_1) = \frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \frac{1}{6} \cdot 3 + \frac{1}{6} \cdot 4 + \frac{1}{6} \cdot 5 + \frac{1}{6} \cdot 6 = \frac{21}{6} = \frac{7}{2} = 3.5,$$

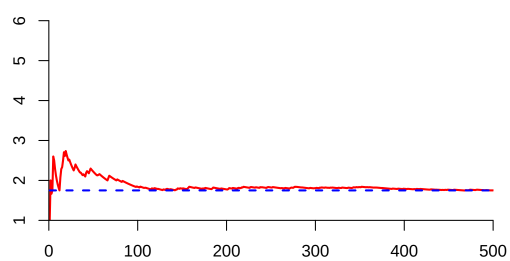
since each of the sides has an equal probability of appearing. With die d_2, however, we cannot say that the average value we would expect to roll is also 3.5. Even though die d_2 has the same number and value of sides as die d_1, we expect to roll a 1 with die d_2 far more often than any other number. Therefore, we must weight the outcome of rolling a 1 accordingly. Doing so gives us an average value of

$$\mathrm{E}(X_2) = 0.75 \cdot 1 + 0.05 \cdot 2 + 0.05 \cdot 3 + 0.05 \cdot 4 + 0.05 \cdot 5 + 0.05 \cdot 6 = \frac{7}{4} = 1.75,$$

which is a much more reasonable average value when we consider that the number of 1s we roll with die d_2 will skew the average closer to 1.

As an illustration of expectation, consider the following figures. The plot on the left shows the cumulative average of 500 randomized rolls of the fair die d_1, and the plot on the right shows the cumulative average of 500 randomized rolls of the biased die d_2. In both plots, the expectation of the experiment is denoted by a dashed line, and we can see that each of the cumulative averages converge to the expectation as we perform more trials.





By Definition 15, the expectation of a random variable is a weighted sum over each outcome in a sample space S. This definition works well for small sample spaces, like for coin flips or die rolls. For larger sample spaces, however, it can be inconvenient to deal with a huge sum consisting of many terms. Luckily, there is an alternative formula to calculate expectation that is written in terms of the values produced by the random variable X rather than the individual outcomes in S.

Theorem 17. *Let (S, P) be a sample space and let X be a real-valued random variable defined on S. Then,*

$$\mathrm{E}(X) = \sum_{a \in \mathbb{R}} a \cdot P(X = a).$$

Proof. Omitted (see Proposition 34.4 on p. 237 of the textbook). $\square$

Note that, although the Proposition above seems to be an infinite sum over all $a \in \mathbb{R}$, in reality since S is a finite set there are only a finite number of values that $X(s)$ can attain. All other values of a that fall outside of the image of X (remember, X is a function) have $P(X = a) = 0$ so we do not need to worry about including them in the sum. Thus, what we really have is a finite sum over the real numbers a for which $P(X = a) > 0$.

Example 18. Recall the spinner from Example 4. Consider a game where you receive \$10 for spinning an odd number and \$20 for spinning an even number. Let X be the payout from this game. What is $\mathrm{E}(X)$? In other words, how much money do we expect to receive per spin if we play this game many times?

By Definition 15, this is

$$\begin{aligned}
\mathrm{E}(X) &= \sum_{s \in S} X(s)P(s) \\
&= X(1)P(1) + X(2)P(2) + X(3)P(3) + X(4)P(4) + X(5)P(5) + X(6)P(6) \\
&= 10 \cdot \frac{1}{2} + 20 \cdot \frac{1}{10} + 10 \cdot \frac{1}{10} + 20 \cdot \frac{1}{10} + 10 \cdot \frac{1}{10} + 20 \cdot \frac{1}{10} = 13
\end{aligned}$$

Alternatively, by Proposition 17, we get

$$\begin{aligned}
\mathrm{E}(X) &= \sum_{a \in \mathbb{R}} a \cdot P(X = a) \\
&= 10 \cdot P(X = 10) + 20 \cdot P(X = 20) \\
&= 10 \cdot \frac{7}{10} + 20 \cdot \frac{3}{10} = 7 + 6 = 13
\end{aligned}$$

So, if we play the game repeatedly, we expect to receive an average of \$13 per spin.

Intuitively this makes sense, since the number 1 covers half the area of the spinner, and it is an odd number, so we expect the average earnings per spin to be closer to \$10 than to \$20.

3.2 Linearity of Expectation

Consider the multiplication of a random variable by a constant. Recall that the expectation of a random variable is essentially just a sum. If we introduce a multiplicative constant into the sum, we can imagine this process either as scaling each value taken by the random variable by some constant factor or, by the following property, as scaling the expectation itself by the same constant factor.

Proposition 19. *Let X be a random variable on some sample space (S, P) and let c be a real number. Then,*

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

Example 20. Recall X_1 from Example 16, which maps the outcomes of rolling a die to the numeric value of the rolled die. We already know that $\mathrm{E}(X) = 3.5$.

Now, suppose a game where you receive $2X$ dollars each time you roll the die. By Proposition 19, we have

$$\mathrm{E}(2X) = 2 \cdot \mathrm{E}(X) = 2 \cdot 3.5 = 7.$$

So, the average earnings you would expect to receive per roll of the die is \$7.

Given two random variables X and Y, what can we say about the expectation of their sum, $X + Y$?

Proposition 21. *Suppose X and Y are real-valued variables defined on a sample space (S, P). Then,*

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

Proof. Omitted (see Proposition 34.7 on p. 239 of the textbook). □

Example 22. Recall X_1 and X_2 from Example 16, which map the outcomes of rolling a die to the numeric value of the rolled die. X_1 corresponds to a fair die d_1, and X_2 corresponds to a biased die d_2 where $P(X_2 =$ ⚅$) = 0.75$ and $P(X_2 =$ ⚄$) = P(X_2 =$ ⚃$) = P(X_2 =$ ⚂$) = P(X_2 =$ ⚁$) = P(X_2 =$ ⚀$) = 0.05$.

Suppose that you roll both dice and take the sum of the outcomes of the dice. What is the expected value of the sum of both dice? We already know that $\mathrm{E}(X_1) = 3.5$ and $\mathrm{E}(X_2) = 1.75$. By Proposition 21, we have

$$\mathrm{E}(X_1 + X_2) = \mathrm{E}(X_1) + \mathrm{E}(X_2) = 3.5 + 1.75 = 5.25.$$

More generally, by combining Propositions 19 and 21 and considering m random variables instead of two random variables, we have the following result.

Corollary 23. *Let $X_1, X_2, \ldots, X_m$ be random variables on some sample space Ω, and let $c_1, c_2, \ldots, c_m$ be real numbers. Then*

$$\mathrm{E}(c_1 X_1 + c_2 X_2 + \cdots + c_m X_m) = c_1 \mathrm{E}(X_1) + c_2 \mathrm{E}(X_2) + \cdots + c_m \mathrm{E}(X_m).$$

Proof. Follows from Propositions 19 and 21. □

3.3 Product of Random Variables

Example 24. A pair of dice are rolled. Let X be the product of the numbers on the two dice. What is the expected value of X?

Let X_1 be number on the first die, and let X_2 be the number on the second die. We can express X as the product of X_1 and X_2. Since we know that $\mathrm{E}(X_1) = \mathrm{E}(X_2) = {}^7\!/_2$, it seems reasonable to guess that

$$\mathrm{E}(X) = \mathrm{E}(X_1)\mathrm{E}(X_2) = ({}^7\!/_2)^2.$$

We can double-check our answer by computing $\sum_{s \in S} X(s)P(s)$ (Definition 15), i.e., by computing the product of the numbers on the two dice for each possible outcome and multiplying it by its probability of occurrence (${}^1\!/_{36}$, since each occurrence is equally probable) and summing across all outcomes. Alternatively we may use Proposition 17 as is done on pp. 241–242 of the textbook.

Either method will confirm that our guess, $({}^7\!/_2)^2$, is correct. However, while $\mathrm{E}(X_1 X_2) = \mathrm{E}(X_1)\mathrm{E}(X_2)$ holds true for this example, it is not true in general.

Example 25. A fair coin is tossed twice. Let X_H be the number of Ⓗ produced and let X_T be the number of Ⓣ produced. Let $Z = X_H X_T$. What is $\mathrm{E}(Z)$?

We know that $\mathrm{E}(X_H) = \mathrm{E}(X_T) = 1$. Note that

$$Z(\text{ⒽⒽ}) = 0, \quad Z(\text{ⓉⓉ}) = 0, \quad Z(\text{ⒽⓉ}) = 1, \text{ and } Z(\text{ⓉⒽ}) = 1.$$

So you may see already that $\mathrm{E}(Z) \neq 1$ and thus $\mathrm{E}(Z) \neq \mathrm{E}(X_H)\mathrm{E}(X_T)$. We can instead compute $\mathrm{E}(Z)$ as follows, by Proposition 17:

$$\begin{aligned}
\mathrm{E}(Z) &= \sum_{a \in \mathbb{R}} a P(Z = a) \\
&= 0 \cdot P(Z = 0) + 1 \cdot P(Z = 1) \\
&= 0 + 1 \cdot \frac{2}{4} \\
&= \frac{1}{2}.
\end{aligned}$$

Notice in the two prior examples that X_1 and X_2 are independent random variables, whereas X_H and X_T are dependent.

Theorem 26. *Let X and Y be independent, real-valued random variables defined on a sample space (S, P). Then,*

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

Omitted (see Theorem 34.14 on p. 242 of the textbook).

Remark. Be careful: the converse of Theorem 26 does not necessarily hold. Even if we know that $\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y)$, we cannot conclude that X and Y are independent.

3.4 Centrality and Variance

The "centrality" of X is a measure of how "spread out" values of X are relative to the expectation of X. What does it mean for values to be "spread out"? To be sure, we need to formalize the notion of "spreading out", but we will begin our discussion by considering an example.

Consider again the sample space (S_{2D}, P_{2D}) for rolling a pair of fair dice. Let X assign each outcome $s \in S$ to the sum of the values on both dice. By the multiplication principle, there are $6 \times 6 = 36$ outcomes in the sample space S_{2D}. However, X has only 11 possible values (the smallest sum is $1 + 1 = 2$ and the largest sum is $6 + 6 = 12$), so by the pigeonhole principle, at least one value of X must correspond to more than one outcome $s \in S$. Indeed, we have the following sums, corresponding outcomes, and probabilities:

2: ⚀⚀	$P(X = 2) = 1/36$
3: ⚀⚁ ⚁⚀	$P(X = 3) = 2/36$
4: ⚀⚂ ⚁⚁ ⚂⚀	$P(X = 4) = 3/36$
5: ⚀⚃ ⚁⚂ ⚂⚁ ⚃⚀	$P(X = 5) = 4/36$
6: ⚀⚄ ⚁⚃ ⚂⚂ ⚃⚁ ⚄⚀	$P(X = 6) = 5/36$
7: ⚀⚅ ⚁⚄ ⚂⚃ ⚃⚂ ⚄⚁ ⚅⚀	$P(X = 7) = 6/36$
8: ⚁⚅ ⚂⚄ ⚃⚃ ⚄⚂ ⚅⚁	$P(X = 8) = 5/36$
9: ⚂⚅ ⚃⚄ ⚄⚃ ⚅⚂	$P(X = 9) = 4/36$
10: ⚃⚅ ⚄⚄ ⚅⚃	$P(X = 10) = 3/36$
11: ⚄⚅ ⚅⚄	$P(X = 11) = 2/36$
12: ⚅⚅	$P(X = 12) = 1/36$

Observe that the outcomes are more concentrated around the value 7, which is the expected value of the sum obtained by rolling two dice. In other words, the probability of rolling two dice and obtaining the value 7, or a value close to 7, is higher than the probability of rolling a value closer to either extreme (2 or 12). So, we can see that X is not very "spread out".

Now, consider the sample space (S_D, P_D) for rolling a single fair die. Let Y assign each outcome $s \in S$ to the value shown on the face of the die. The expected value of Y is 3.5, but each value of Y is equally probable—i.e., the probability of rolling a number closer to the expected value is the same as the probability of rolling a number closer to either extreme (1 or 6). So, we can see that Y is more "spread out" in comparison to X.

The following definition provides a mathematical measure of a random variable's "spread", called **variance**.

Definition 27 (Variance)**.** Let X be a real-valued random variable on a sample space (S, P). Let $\mu = \mathrm{E}(X)$. The **variance** of X is

$$\begin{aligned}\mathrm{Var}(X) &= \mathrm{E}\left[(X-\mu)^2\right] \\ &= \sum_{s \in S} \mathrm{P}(s)\,(X(s)-\mu)^2.\end{aligned}$$

Observe that if an outcome s is farther from the expected value $\mathrm{E}(X)$, this will cause $(X(s)-\mu)^2$ to be larger. If the outcome s also has a high probability of occurring, $\mathrm{P}(s)$ will also be larger. So, we can see intuitively that a random variable X will have a higher variance if outcomes farther from the average occur with higher probability.

Immediately from our definition, we get a handy formulation of variance exclusively in terms of expectation.

Theorem 28. *Let X be a random variable. Then*

$$\mathrm{Var}(X) = \mathrm{E}[X^2] - \mathrm{E}[X]^2.$$

Proof. Let $\mu = \mathrm{E}[X]$. By Definition 27, we have that

$$\begin{aligned}\mathrm{Var}(X) &= \mathrm{E}[(X-\mu)^2] \\ &= \mathrm{E}[X^2 - 2X\mu + \mu^2] \\ &= \mathrm{E}[X^2] - 2\mu\mathrm{E}[X] + E[\mu^2] \\ &= \mathrm{E}[X^2] - 2\mu^2 + \mu^2 \\ &= \mathrm{E}[X^2] - \mu^2 \\ &= \mathrm{E}[X^2] - \mathrm{E}[X]^2.\end{aligned}$$

□

Example 29. Consider an experiment where we roll one die. Let X be a random variable mapping each outcome to the number showing on the die. What is $\mathrm{Var}(X)$?

Let us first calculate $\mathrm{Var}(X)$ using Definition 27. So, let $\mu = \mathrm{E}(X) = \frac{7}{2}$. Then,

$$\begin{aligned}\mathrm{Var}(X) &= \mathrm{E}\left[(X-\mu)^2\right] = \mathrm{E}\left[(X-{}^7\!/\!_2)^2\right] \\ &= (1-{}^7\!/\!_2)^2\cdot{}^1\!/\!_6 + (2-{}^7\!/\!_2)^2\cdot{}^1\!/\!_6 + (3-{}^7\!/\!_2)^2\cdot{}^1\!/\!_6 + (4-{}^7\!/\!_2)^2\cdot{}^1\!/\!_6 + (5-{}^7\!/\!_2)^2\cdot{}^1\!/\!_6 + (6-{}^7\!/\!_2)^2\cdot{}^1\!/\!_6 \\ &= \frac{25}{24} + \frac{3}{8} + \frac{1}{24} + \frac{1}{24} + \frac{3}{8} + \frac{25}{24} \\ &= \frac{35}{12} \approx 2.9167\end{aligned}$$

Now, let us calculate $\mathrm{Var}(X)$ using Theorem 28: $\mathrm{Var}(X) = \mathrm{E}[X^2] - \mathrm{E}[X]^2$.

Note that $\mathrm{E}[X]^2 = (7/2)^2 = 49/4$. We calculate $\mathrm{E}[X^2]$ as follows.

$$\begin{aligned}\mathrm{E}[X^2] &= 1^2 \cdot 1/6 + 2^2 \cdot 1/6 + 3^2 \cdot 1/6 + 4^2 \cdot 1/6 + 5^2 \cdot 1/6 + 6^2 \cdot 1/6 \\ &= \frac{1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2}{6} = \frac{91}{6}.\end{aligned}$$

So, we have

$$\begin{aligned}\mathrm{Var}(X) &= \mathrm{E}[X^2] - \mathrm{E}[X]^2 \\ &= \frac{91}{6} - \frac{49}{4} = \frac{35}{12}.\end{aligned}$$

We can see that this agrees with the answer we obtained using Definition 27.

Now, let us assume that we rig the die so that it never rolls a ⚀ or a ⚅, but each of the remaining four outcomes ⚁, ⚂, ⚃, and ⚄ occur with equal probability $1/4$. Let Y be a random variable mapping each outcome to the number showing on the die. Notice that $\mathrm{E}(Y)$ is $\frac{7}{2}$, which is the same as $\mathrm{E}(X)$. However, the variance will be different. Let us compute it using Definition 27:

$$\begin{aligned}\mathrm{Var}(Y) &= \mathrm{E}\left[(Y-\mu)^2\right] = \mathrm{E}\left[(Y - 7/2)^2\right] \\ &= (2 - 7/2)^2 \cdot 1/4 + (3 - 7/2)^2 \cdot 1/4 + (4 - 7/2)^2 \cdot 1/4 + (5 - 7/2)^2 \cdot 1/4 \\ &= \frac{9}{16} + \frac{1}{16} + \frac{1}{16} + \frac{3}{8} \\ &= \frac{5}{4} = 1.25\end{aligned}$$

Notice that the variance of Y is less than the variance of X. This makes sense, because after having removed two extreme outcomes ⚀ and ⚅ from the experiment, the outcomes that remain are the ones that have a value closer to the average value of 3.5, so Y is less "spread out" than X.

Exercise

Calculate $\mathrm{Var}(X)$ for the random variable X on the sample space (S_{2D}, P_{2D}) given in the opening example of this section. Notice that the answer, $\frac{35}{6}$, is double compared to the variance of the experiment with a single die computed in the previous example. This might seem strange, since we saw that the outcomes for the two-die experiment are more closely clustered around the average value, whereas the outcomes for the one-die experiment are all equally likely. This shows us that it can be misleading to compare $\mathrm{Var}(X)$ and $\mathrm{Var}(Y)$ if $\mathrm{E}(X) \neq \mathrm{E}(Y)$. There are "normalized" measures (which are outside the scope of this course) that allow a more "fair" way to compare the "spread" of random variables with different average values.