Chapter 1

Revision of discrete probability theory

1.1 Outline

In the chapter we review some of the basic definitions and results of probability theory that we are going to use throughout the course. You will have covered most of this material in MM204 and MM304, although the emphasis may be a little different.

1.2 Set theory notation

A *set* is a collection of *elements*. The set of no elements is the *empty set* $\emptyset$. Finite non-empty sets can be listed like $S = \{a_1, \ldots, a_n\}$. If a set S contains an element a, we write $a \in S$. A set R is a *subset* of a set S, written $R \subseteq S$, if every $a \in R$ also satisfies $a \in S$. For two sets S and T, their *intersection* is $S \cap T$, the set of elements that are in *both* A and B, and their *union* is $S \cup T$, the set of elements in at least one of S or T. For two sets S and T, 'S minus T' is the set $S \setminus T = \{a \in S : a \notin T\}$, the set of elements that are in S but not in T.

Note that $S \cap \emptyset = \emptyset$, $S \cup \emptyset = S$, and $S \setminus \emptyset = S$.

1.3 Discrete probability spaces

Suppose we perform an experiment that gives a random outcome. Although we don't know what the outcome will be, suppose we do know all of the possible outcomes it could be. Let Ω denote the set of all possible outcomes: the *sample space*. To start with, we take Ω to be *discrete*, which means it is *finite* or *countably infinite*. This means that we can write Ω as a (possibly infinite) list:

$$\Omega = \{\omega_1, \omega_2, \omega_3, \ldots\},$$

where $\omega_1, \omega_2, \omega_3, \ldots$ are the possible outcomes to our experiment.

Example. Rolling an ordinary die. $\Omega = \{1, 2, 3, 4, 5, 6\}$. ■

A set $A \subseteq \Omega$ (i.e., a subset of Ω) is called an *event*. It is events that we are interested in. For example, $\{4, 5, 6\}$ is the event that the die scores at least 4; $\{2, 4, 6\}$ is the event that the die is even. The collection of events associated with Ω includes Ω itself (the certain event!) and $\emptyset$ (the impossible event!).

Given events $A, B \subseteq \Omega$, we can build new events using the operations of set theory:

- $A \cup B$ ("A or B"), the event that A happens, or B happens, or both.

- $A \cap B$ ("A and B"), the event that A and B both happen.

Two events A and B are called *disjoint* or *mutually exclusive* if $A \cap B = \emptyset$.

Example. If $A = \{4,5,6\}$, $B = \{2,4,6\}$, and $C = \{1,3\}$ then $A \cup B = \{2,4,5,6\}$ (the score is even or at least 4) and $A \cap B = \{4,6\}$ (the score is even and at least 4). $A \cap C = \emptyset$ so A and C are disjoint; so are B and C. ■

We want to assign probabilities to events.

Definition 1.1. *Let Ω be a non-empty discrete sample space. A function $\mathbb{P}$ that gives a value $\mathbb{P}(A) \in [0,1]$ for every subset $A \subseteq \Omega$ is called a* probability measure *on Ω if:*

(P1) $\mathbb{P}(\emptyset) = 0$ and $\mathbb{P}(\Omega) = 1$;

(P2) For any $A_1, A_2, \ldots$ a collection of disjoint *subsets of Ω,*

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

Given Ω and a probability measure $\mathbb{P}$, we call $(\Omega, \mathbb{P})$ a discrete probability space*.*

For an event $A \subseteq \Omega$, we define its *complement*, denoted A^c (or sometimes $\bar{A}$) and read "not A", to be $A^c := \Omega \setminus A = \{\omega \in \Omega : \omega \notin A\}$. Notice that $(A^c)^c = A$, $A \cap A^c = \emptyset$ and $A \cup A^c = \Omega$.

Example. In the die example, $\Omega = \{1, 2, \ldots, 6\}$; it is natural to suppose that the die is *fair*, so each outcome is equally likely. Then for $A \subseteq \Omega$, $\mathbb{P}(A)$ is $\frac{1}{6}$ times the number of outcomes in A. So $\mathbb{P}(\{2,4,6\}) = \frac{3}{6} = \frac{1}{2}$ and $\mathbb{P}(\{4,5,6\}) = \frac{1}{2}$ too. ■

Note: for non-discrete sample spaces, we may not be able to assign probabilities to *all* subsets in a sensible way, and so smaller collections of events are required.

Theorem 1.2 (Properties of probability)**.** *Let $(\Omega, \mathbb{P})$ be a discrete probability space. Then:*

(i) For $A \subseteq \Omega$, $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$;

(ii) If $A, B \subseteq \Omega$ and $A \subseteq B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$ [monotonicity];

(iii) If $A, B \subseteq \Omega$, then

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

In particular, if $A \cap B = \emptyset$, $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.

Proof. Exercise. □

Example. Back to the die example, if $A = \{4,5,6\}$ and $B = \{2,4,6\}$, $A \cup B = \{2,4,5,6\}$, $A \cap B = \{4,6\}$, and (iii) says that

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) = \frac{3}{6} + \frac{3}{6} - \frac{2}{6} = \frac{4}{6}.$$

■

Definition 1.3 (Conditional probability)**.** *If A and B are events with $\mathbb{P}(B) > 0$ then the* conditional probability *$\mathbb{P}(A \mid B)$ of A given B is defined by*

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

Suppose that we know that an event B has occurred. Then $\mathbb{P}(A \mid B)$ is the probability that A occurs given B has occurred.

Theorem 1.4. *For any $B \subseteq \Omega$ with $\mathbb{P}(B) > 0$, $\mathbb{P}(\cdot \mid B)$ is a probability measure on Ω.*

Proof. Exercise. □

So $\mathbb{P}(\cdot \mid B)$ behaves just like $\mathbb{P}(\cdot)$. For many calculations, it is more useful to write

$$\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B)\mathbb{P}(B). \tag{1.1}$$

Example. Suppose an urn contains 2 blue and 3 red balls. Two balls are drawn *without* replacement. What is the probability that the first ball is blue and the second is red? Let B be the event that the first ball is blue, and A be the event that the second ball is red. We have $\mathbb{P}(B) = 2/5$ and $\mathbb{P}(A \mid B) = 3/4$. Hence

$$\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B)\mathbb{P}(B) = \frac{2}{5} \cdot \frac{3}{4} = \frac{3}{10}.$$

■

More generally, we have the following result.

Theorem 1.5 (Multiplication rule for conditional probabilities)**.** *Let $A_1, \ldots, A_n$ be finitely many events. Suppose that $\mathbb{P}(A_1 \cap \cdots \cap A_{n-1}) > 0$. Then*

$$\mathbb{P}(A_1 \cap \cdots \cap A_n) = \mathbb{P}(A_1) \cdot \mathbb{P}(A_2 \mid A_1) \cdot \mathbb{P}(A_3 \mid A_1 \cap A_2) \cdots \mathbb{P}(A_n \mid A_1 \cap \cdots \cap A_{n-1}). \tag{1.2}$$

To see what's going on, consider three events, A_1, A_2, A_3, with $\mathbb{P}(A_1 \cap A_2) > 0$. Then, since $A_1 \cap A_2 \subseteq A_1$, we know that $\mathbb{P}(A_1) \geq \mathbb{P}(A_1 \cap A_2) > 0$ by the monotonicity of $\mathbb{P}$. Then,

$$\mathbb{P}(A_1 \cap A_2 \cap A_3) = \mathbb{P}(A_1 \cap A_2)\mathbb{P}(A_3 \mid A_1 \cap A_2) = \mathbb{P}(A_1)\mathbb{P}(A_2 \mid A_1)\mathbb{P}(A_3|A_1 \cap A_2),$$

by two applications of (1.1). The idea behind the proof of the theorem should now be clear.

Example. Suppose you are at a party with n people (in total). We use the multiplication rule to calculate the probability that all the n people have a different birthday. Call this event D_n. We ignore the 29^{th} of February and suppose that a year has 365 days. List the guests in some order as $1, 2, \ldots, n$. Let A_i be the event that the i^{th} guest does not share a birthday with any of guests $1, 2, \ldots, i-1$. Then $D_n = \cap_{i=1}^n A_i$ and

$$\mathbb{P}(A_1) = 1; \quad \mathbb{P}(A_2 \mid A_1) = \frac{364}{365}; \quad \mathbb{P}(A_3 \mid A_1 \cap A_2) = \frac{363}{365},$$

up to, for $n \leq 365$,

$$\mathbb{P}(A_n \mid A_1 \cap \cdots \cap A_{n-1}) = \frac{365 - n + 1}{365}.$$

Thus we obtain, for $1 \leq n \leq 365$

$$\mathbb{P}(A_1 \cap \cdots \cap A_n) = \prod_{i=1}^{n-1} \frac{365 - i}{365} = \frac{364!}{(365-n)!365^{n-1}}.$$

It turns out that $\mathbb{P}(D_{22}) \approx 0.5243$ but $\mathbb{P}(D_{23}) \approx 0.4927$, so as soon as there are at least 23 people, there's a better than 50% chance that two of them share a birthday! ■

Definition 1.6 (Partition)**.** *A countable collection of events $E_1, E_2, \ldots$ is called a* partition *of* Ω *if:*

(i) For all i, $E_i \subseteq \Omega$ and $\mathbb{P}(E_i) > 0$;

(ii) For $i \neq j$, $E_i \cap E_j = \emptyset$ (the events are disjoint);

(iii) $\bigcup_i E_i = \Omega$ (the events fill the sample space).

In many cases, a useful partition consists of just two events, A and A^c.

Although the following theorem follows directly from the definition, we shall see that is is very useful.

Theorem 1.7 (Law of total probability)**.** *Let $E_1, E_2, \ldots$ be a partition of Ω. Then for all $A \subseteq \Omega$,*

$$\mathbb{P}(A) = \sum_i \mathbb{P}(E_i)\mathbb{P}(A \mid E_i).$$

Proof. Given a partition $(E_i, i \in I)$ of Ω, we have

$$A = A \cap \Omega = A \cap \left(\bigcup_{i \in I} E_i \right) = \bigcup_{i \in I} (A \cap E_i),$$

which is the distributive law for sets. Since the $(E_i, i \in I)$ are pairwise disjoint, so are the $(A \cap E_i, i \in I)$. Hence,

$$\mathbb{P}(A) = \mathbb{P}\left(\bigcup_{i \in I} (A \cap E_i) \right) = \sum_{i \in I} \mathbb{P}(A \cap E_i) = \sum_{i \in I} \mathbb{P}(A \mid E_i)\mathbb{P}(E_i),$$

using the definition of conditional probability (since $\mathbb{P}(E_i) > 0$). □

Example. Again consider an urn with two blue balls and three red balls, and suppose that two balls are drawn without replacement. What is the probability of the event E that both balls have the same colour?

Let B, A be the event that the first ball is blue, red respectively. Clearly, (A, B) is a partition of Ω. Then, by the law of total probability,

$$\mathbb{P}(E) = \mathbb{P}(B)\mathbb{P}(E \mid B) + \mathbb{P}(A)\mathbb{P}(E \mid A) = \frac{2}{5} \cdot \frac{1}{4} + \frac{3}{5} \cdot \frac{2}{4} = \frac{2}{5}.$$

■

Theorem 1.8 (Bayes' formula)**.** *Let $E_1, E_2, \ldots$ be a partition of Ω. Then for all $A \subseteq \Omega$ with $\mathbb{P}(A) > 0$,*

$$\mathbb{P}(E_n \mid A) = \frac{\mathbb{P}(A \mid E_n)\mathbb{P}(E_n)}{\sum_i \mathbb{P}(A \mid E_i)\mathbb{P}(E_i)}.$$

Proof. By the law of total probability, the denominator is just $\mathbb{P}(A)$. The numerator is $\mathbb{P}(A \cap E_n)$ by the definition of conditional probability. Hence the theorem follows, again by the definition of conditional probability. □

Example. Bob the cat has gone missing. It is presumed Bob is equally likely to be in any one of three places: the arcade, the butcher's, and the chip shop. The probability that a search of the arcade would turn up Bob, if Bob was actually there, is 4/5. The corresponding probabilities for the butcher's and chip shop are 1/2 and 3/4 respectively. What is the (conditional) probability that Bob is in the butcher's given that searches of the arcade and the chip shop have been unsuccessful?

By Bayes' formula

$$\mathbb{P}(B \mid U_A \cap U_C) = \frac{\mathbb{P}(U_A \cap U_C \mid B)\mathbb{P}(B)}{\mathbb{P}(U_A \cap U_C)} = \frac{1}{3\mathbb{P}(U_A \cap U_C)}.$$

By the law of total probability (LOTP),

$$\begin{aligned}\mathbb{P}(U_A \cap U_C) &= (U_A \cap U_C \mid A)\mathbb{P}(A) + (U_A \cap U_C \mid B)\mathbb{P}(B) + (U_A \cap U_C \mid C)\mathbb{P}(C) \\ &= \frac{1}{3}\left(\frac{1}{5} + 1 + \frac{1}{4}\right) = \frac{29}{60}.\end{aligned}$$

So $\mathbb{P}(B \mid U_A \cap U_C) = 20/29 \approx 0.69$. ■

Definition 1.9 (Independent events)**.** *A countable collection* $(A_i, i \in I)$ *of events is called* independent *if, for every* finite *subset* $J \subseteq I$*,*

$$\mathbb{P}\Big(\bigcap_{j \in J} A_j\Big) = \prod_{j \in J} \mathbb{P}(A_j).$$

In particular, two events A and B are independent if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

Example. For a fair die, if $A = \{4, 5, 6\}$ and $B = \{2, 4, 6\}$, then $A \cap B = \{4, 6\}$. So $\mathbb{P}(A \cap B) = \frac{2}{6} = \frac{1}{3}$ which is not the same as $\mathbb{P}(A) \times \mathbb{P}(B) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$. So A and B are *not* independent. ■

If events $(A_i, i \in I)$ are independent, then they are also *pairwise independent*, i.e., for $i, j \in I$ with $i \neq j$, $\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j)$. However, the converse is not true in general, as the following example shows.

Example. Let $\Omega = \{1, 2, 3, 4\}$ and $\mathbb{P}(i) = 1/4$ for $i \in \Omega$. Define the events $A = \{1, 2\}$, $B = \{1, 3\}$ and $C = \{2, 3\}$. Then (check!) A, B, C are pairwise independent but *not* independent. ■

Note that if A and B are independent and $\mathbb{P}(B) > 0$, then $\mathbb{P}(A \mid B) = \mathbb{P}(A)$.

1.4 Random variables and expectation

Let $(\Omega, \mathbb{P})$ be a discrete probability space. A function $X : \Omega \to \mathbf{R}$ is a *random variable*. So each $\omega \in \Omega$ is mapped to a real number $X(\omega)$. The set of possible values for X is $X(\Omega) = \{X(\omega) : \omega \in \Omega\} \subset \mathbf{R}$. Notice that since Ω is discrete, $X(\Omega)$ must be also.

If X and Y are two random variables on $(\Omega, \mathbb{P})$, then $X + Y$, XY, etc, are also random variables. E.g., $(X + Y)(\omega) = X(\omega) + Y(\omega)$.

Definition 1.10. *The* distribution *of a discrete r.v.* X *is given by* $\mathbb{P}(X = x)$ *for all* $x \in X(\Omega)$*.*

Definition 1.11. *The* distribution function *of* X *is* $F : \mathbf{R} \to [0, 1]$ *given by* $F(x) = \mathbb{P}(X \leq x)$*.*

Example. (Binomial distribution.) Let n be a positive integer and $p \in [0, 1]$. If for $k \in \{0, 1, \ldots, n\}$, $\mathbb{P}(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$, X is a binomial random variable with parameters n, p. We write $X \sim \mathrm{Bin}(n, p)$. The binomial distribution has the following interpretation: Perform

n independent "trials" (e.g., coin tosses) each with probability p of "success" (e.g., "heads"), and count the total number of successes. ■

Example. (Poisson distribution.) Let $\lambda > 0$ and $p_k := \mathrm{e}^{-\lambda}\frac{\lambda^k}{k!}$ for $k = 0, 1, 2, \ldots$. If $\mathbb{P}(X = k) = p_k$, X is a Poisson random variable with parameter λ. We write $X \sim \mathrm{Po}(\lambda)$. ■

Definition 1.12 (Expectation of a discrete r.v.)**.** *Let X be a discrete random variable. The* expectation, expected value, *or* mean *of X is defined by:*

$$\mathbb{E}(X) = \sum_{x \in X(\Omega)} x\mathbb{P}(X = x),$$

provided the sum is finite.

Example. Rolling a fair die. The score X is the *discrete uniform* distribution on $S = \{1, 2, \ldots, 6\}$, i.e., $\mathbb{P}(X = x) = \frac{1}{6}$ for $x \in S$. Then $F(x) = \mathbb{P}(X \leq x) = \frac{x}{6}$ and

$$\mathbb{E}(X) = \sum_{x \in S} x\mathbb{P}(X = x) = \frac{1}{6}(1 + 2 + \cdots + 6) = \frac{7}{2}.$$

■

Example. Let A be an event. Let $\mathbf{1}_A$ denote the *indicator random variable* of A, that is, $\mathbf{1}_A : \Omega \to \{0, 1\}$ given by

$$\mathbf{1}_A(\omega) := \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A \end{cases}$$

So $\mathbf{1}_A$ is 1 if A happens and 0 if not. Then

$$\mathbb{E}(\mathbf{1}_A) = 1 \cdot \mathbb{P}(\mathbf{1}_A = 1) + 0 \cdot \mathbb{P}(\mathbf{1}_A = 0) = \mathbb{P}(\mathbf{1}_A = 1) = \mathbb{P}(A).$$

■

Theorem 1.13 (Properties of expectation)**.** *For X and Y random variables with well defined expectations and $a, b \in \mathbf{R}$,*

(a) Linearity: $\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$.

(b) Monotonicity: Let $X \leq Y$, i.e. $X(\omega) \leq Y(\omega)$ for all $\omega \in \Omega$. Then $\mathbb{E}(X) \leq \mathbb{E}(Y)$.

(c) Triangle inequality: $|\mathbb{E}(X)| \leq \mathbb{E}(|X|)$.

(d) Law of the Unconscious Statistician: Let $h : X(\Omega) \to \mathbf{R}$. Then

$$\mathbb{E}(h(X)) = \sum_{x \in X(\Omega)} h(x)\mathbb{P}(X = x).$$

Proof. Exercise. □

Definition 1.14. *We define the* variance *of a random variable X as*

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

Example. Let Y be a Bernoulli random variable with parameter $p \in [0,1]$, so that Y takes values in $S = \{0,1\}$ with $\mathbb{P}(Y=1) = p$ and $\mathbb{P}(Y=0) = 1-p$. Then

$$\mathbb{E}(Y) = 0 \times \mathbb{P}(Y=0) + 1 \times \mathbb{P}(Y=1) = p,$$

and

$$\mathbb{E}(Y^2) = 0^2 \times \mathbb{P}(Y=0) + 1^2 \times \mathbb{P}(Y=1) = p,$$

so $\mathrm{Var}(Y) = p - p^2 = p(1-p)$. ■

Definition 1.15 (Independence of random variables)**.** *Let* $(\Omega, \mathbb{P})$ *be a discrete probability space. A family* $(X_i, i \in I)$ *of random variables is called* independent *if for any finite subset* $J \subseteq I$ *and all* $x_j \in X_j(\Omega)$,

$$\mathbb{P}\left(\bigcap_{j \in J} \{X_j = x_j\}\right) = \prod_{j \in J} \mathbb{P}(X_j = x_j).$$

In particular, random variables $X_1, \ldots, X_n$ are independent if:

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \mathbb{P}(X_1 = x_1) \cdots \mathbb{P}(X_n = x_n),$$

for all $x_1, \ldots, x_n$. $\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)$ is called the *joint distribution* of $X_1, \ldots, X_n$.

Theorem 1.16 (Independence means multiply)**.** *Let X and Y be independent random variables on probability space* $(\Omega, \mathbb{P})$*. Then* $\mathbb{E}(XY) = \mathbb{E}(X) \cdot \mathbb{E}(Y)$*.*

Theorem 1.17. *Let X and Y be independent random variables on probability space* $(\Omega, \mathbb{P})$ *and let* $a, b \in \mathbf{R}$*. Then*

$$\mathrm{Var}(aX + bY) = a^2 \mathrm{Var}(X) + b^2 \mathrm{Var}(Y).$$

In particular, if X and Y are *independent*, $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

Example. Suppose $X \sim \mathrm{Bin}(n,p)$. What are $\mathbb{E}(X)$ and $\mathrm{Var}(X)$? Note X can be written as

$$X = \sum_{i=1}^{n} Y_i,$$

where Y_i are independent, identically distributed (i.i.d.) Bernoulli random variables taking values 1 (success) and 0 (failure) with probabilities p and $1-p$ respectively. Recall that $\mathbb{E}(Y_i) = p$ and $\mathrm{Var}(Y_i) = p(1-p)$. Then

$$\mathbb{E}(X) = \mathbb{E} \sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \mathbb{E}(Y_i) = np,$$

by linearity of expectation. Also, by independence, $\mathrm{Var}(X) = \sum_{i=1}^{n} \mathrm{Var}(Y_i) = np(1-p)$. ■

Example. Minimum and maximum of probability distributions. Let X, Y be the scores on two independent rolls of a fair die, and let $Z = \max\{X, Y\}$ be the maximum of the two. What is the distribution of Z and what is $\mathbb{E}(Z)$?

We have that $Z \leq x$ if and only if $X \leq x$ *and* $Y \leq x$. So

$$\mathbb{P}(Z \leq x) = \mathbb{P}(\{X \leq x\} \cap \{Y \leq x\}) = \mathbb{P}(X \leq x) \cdot \mathbb{P}(Y \leq x),$$

by independence. So

$$\mathbb{P}(Z \leq x) = \frac{x}{6} \cdot \frac{x}{6} = \frac{x^2}{36}.$$

To work out $\mathbb{P}(Z = x)$ we can use the fact that

$$\mathbb{P}(Z = x) = \mathbb{P}(Z \le x) - \mathbb{P}(Z \le x - 1) = \frac{x^2}{36} - \frac{(x-1)^2}{36} = \frac{2x-1}{36}.$$

(You can also try to check this directly, by counting all the possibilities.) So

$$\mathbb{E}(Z) = \sum_{x=1}^{6} \frac{(2x-1)x}{36} = \frac{161}{36} \approx 4.47.$$

■

1.5 Conditional expectation

Definition 1.18 (Conditional expectation with respect to an event)**.** *On a discrete probability space* $(\Omega, \mathbb{P})$ *let* B *be an event with* $\mathbb{P}(B) > 0$ *and let* X *be a random variable. The* conditional expectation *of* X *given* B *is*

$$\mathbb{E}(X \mid B) = \sum_{x \in X(\Omega)} x\mathbb{P}(X = x \mid B).$$

So $\mathbb{E}(X \mid B)$ can be thought of as expectation with respect to the conditional probability measure $\mathbb{P}(\cdot \mid B)$. There is an alternative formula which is often very useful:

Theorem 1.19. *For an event* B *with* $\mathbb{P}(B) > 0$*,*

$$\mathbb{E}(X \mid B) = \frac{\mathbb{E}(X\mathbf{1}_B)}{\mathbb{P}(B)},$$

where $\mathbf{1}_B$ *is the indicator random variable of* B*.*

Proof. The proof is an exercise in interchanging summations. Starting from our definition,

$$\mathbb{E}(X \mid B) = \sum_{x \in X(\Omega)} x\mathbb{P}(X = x \mid B) = \sum_{x \in X(\Omega)} x\frac{\mathbb{P}(\{X = x\} \cap B)}{\mathbb{P}(B)},$$

by the definition of conditional probability. The random variable $\mathbf{1}_B X$ takes values $x \neq 0$ with

$$\mathbb{P}(\mathbf{1}_B X = x) = \sum_{\omega \in \Omega : \omega \in B \cap \{X = x\}} \mathbb{P}(\{\omega\}) = \mathbb{P}(\{X = x\} \cap B),$$

so by the definition of expectation

$$\frac{1}{\mathbb{P}(B)} \sum_{x \in X(\Omega)} x\mathbb{P}(\{X = x\} \cap B) = \frac{1}{\mathbb{P}(B)} \cdot \mathbb{E}(\mathbf{1}_B X).$$

□

Theorem 1.20 (Partition theorem for expectations)**.** *Let* $(E_i, i \in I)$ *be a partition of* Ω*. Then for a random variable* X*,*

$$\mathbb{E}(X) = \sum_{i \in I} \mathbb{E}(X \mid E_i)\mathbb{P}(E_i).$$

Proof. Since $(E_i, i \in I)$ is a partition, we know $\sum_{i\in I} \mathbf{1}_{E_i} = 1$. Hence

$$\mathbb{E}(X) = \mathbb{E}\left(X \sum_{i\in I} \mathbf{1}_{E_i}\right) = \mathbb{E}\left(\sum_{i\in I} X\mathbf{1}_{E_i}\right) = \sum_{i\in I} \mathbb{E}(X\mathbf{1}_{E_i}),$$

by linearity of expectation. By the previous theorem, $\mathbb{E}(X\mathbf{1}_{E_i}) = \mathbb{E}(X \mid E_i)\mathbb{P}(E_i)$. □

Example. We throw a fair die and subsequently toss a fair coin as many times as the number shown on the die. Let X denote the number of heads obtained. What is $\mathbb{E}(X)$?

Let D_i be the event that the score on the die is i. Then $(D_i, i \in \{1, \ldots, 6\})$ is a partition of Ω, with $\mathbb{P}(D_i) = 1/6$ for each i. Given D_i, $X \sim \text{Bin}(i, 1/2)$; in other words, $\mathbb{P}(X = k \mid D_i) = \binom{i}{k}2^{-i}$ for $k \in \{0, \ldots, i\}$. Hence $\mathbb{E}(X \mid D_i) = i/2$, and using the partition theorem

$$\mathbb{E}(X) = \sum_{i=1}^{6} \mathbb{E}(X \mid D_i)\mathbb{P}(D_i) = \frac{1}{6}\sum_{i=1}^{6} \frac{i}{2} = \frac{1}{24} \times 6 \times 7 = \frac{7}{4}.$$

■