Part A

Probability and distribution theory

1

Probability

In this chapter, we introduce some elementary concepts of probability which will be required for the rest of this book. We start by introducing the notion of sets. *Sets* are collections of objects, such as numbers, which are called *elements* of the set. If an element x belongs to set A, we write $x \in A$; otherwise, $x \notin A$. The *empty set* $\varnothing$ contains no elements, while the *universal set* Ω contains all objects of a certain specified type. A set containing a single element is called a *singleton*. The *complement* of set A is the set of all objects in Ω but not included in A. It can be represented by $A^c := \{x : x \notin A\}$, which stands for "$x$ such that $x \notin A$". If a set A includes all the elements of another set B, the latter is called a *subset* of the former, and is denoted by $B \subseteq A$. The two sets may be equal, but if A contains further elements which are not in B, then B is a *proper subset* of A and this is denoted as $B \subset A$.

The *intersection* of two sets A and B is given by the elements belonging to both sets simultaneously, and is defined as $A \cap B := \{x : x \in A \text{ and } x \in B\}$. The collection of elements in set B but not in set A is $B \cap A^c$ and is denoted by $B \backslash A$. Sets A and B are *disjoint* if and only if $A \cap B = \varnothing$. The *union* of two sets is the collection of all elements in either set, and is defined by $A \cup B := \{x : x \in A \text{ or } x \in B\}$. Clearly,

$$A \cap B = B \cap A \quad \text{and} \quad A \cup B = B \cup A,$$

so that intersection and union possess the property of *commutativity*. The *distributive laws*

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

and

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

apply to sets. Finally, an important law of logic, when applied to sets, is *De Morgan's law* which states that

$$(A \cap B)^c = A^c \cup B^c$$

and

$$(A \cup B)^c = A^c \cap B^c;$$

see also Section A.1 in Appendix A. It can be extended to a *countable* collection of sets $A_1, A_2, \dots$ (instead of just A, B), the adjective "countable" meaning that the A_i's can be enumerated by an index such as $i = 1, 2, \dots, \infty$. Notice that countability does not necessarily mean that there is a finite number of A_i's (see the index i); rather it means that the set of natural numbers $\mathbb{N}$ is big enough to count all the A_i's.

Now consider the case when the objects in these sets are outcomes of a *random experiment*, one where chance could lead to a different outcome if the experiment were repeated. Then Ω is called the *sample space*, that is, the collection of all potential outcomes of the experiment.

Consider the most common example of an experiment: tossing a coin where the outcomes are a head (H) or a tail (T). Then the sample space is $\Omega = \{H, T\}$, namely head and tail. If the coin is to be tossed twice, then $\Omega = \{HH, TT, HT, TH\}$ where HT denotes a head followed by a tail.

An *event* A is a subset of Ω. For instance, $A = \{HT\}$ is an event in our last example. We also need to be able to talk about:

- the complement A^c of an event A, to discuss whether the event happens or not;
- the union $A_1 \cup A_2$ of two events A_1, A_2, to describe the event that one or the other (or both) happens;
- hence (by De Morgan's law) also the intersection $B_1 \cap B_2$ of two events B_1, B_2, this being the event where both happen simultaneously.

As will be seen in Chapter 2, the sample space Ω may be too big to have its elements enumerated by $i = 1, 2, \dots$, so let us instead focus on some events of interest, $A_1, A_2, \dots$, and define the following. A *sigma-algebra* (or *sigma-field*) of events, $\mathcal{F}$, is a collection of some events $A_i \subseteq \Omega$ (where $i = 1, 2, \dots, \infty$) that satisfies:

- $\varnothing \in \mathcal{F}$;
- if $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$;
- if $A_1, A_2, \dots \in \mathcal{F}$, then $A_1 \cup A_2 \cup \dots \in \mathcal{F}$.

To illustrate $\mathcal{F}$, recall the simplest case of tossing a coin once, leading to $\Omega = \{H, T\}$. Its largest sigma-algebra is the set of all subsets of Ω, called the *power set* and denoted by 2^Ω in general, and given by $\mathcal{F} = \{\varnothing, \{H\}, \{T\}, \Omega\}$ here; its smallest sigma-algebra is the trivial $\mathcal{F} = \{\varnothing, \Omega\}$, containing the *impossible event* $\varnothing$ (nothing happens) and the *certain event* Ω (either a head or a tail happens). Notice that $\mathcal{F}$ is a set whose elements are themselves sets, and that nonsingleton elements of $\mathcal{F}$ are *composite events*; for example, when tossing a coin twice, $\{\{HT\}, \{TH\}\}$ is the composite event of getting one head and one tail regardless of the order in which this happens.

One can define a measure (or function) on this algebra, called *probability*, satisfying the axioms

$$\Pr(A_i) \geqslant 0 \text{ for } i = 1, 2, \ldots,$$

$$\Pr(A_1 \cup A_2 \cup \cdots) = \Pr(A_1) + \Pr(A_2) + \cdots,$$

$$\Pr(\Omega) = 1,$$

for any sequence of disjoint sets $A_1, A_2, \cdots \in \mathcal{F}$. The second axiom is called *countable additivity*, the property of countability having been built into the definition of $\mathcal{F}$ (on which probability is defined) though it is not always a property of Ω as will be illustrated in Chapter 2. These axioms imply that $\Pr(\varnothing) = 0$ and $\Pr(A_i) \in [0, 1]$.

A *fair coin* is a coin having probability $\frac{1}{2}$ for each outcome. Typically, it is also implicitly assumed that the coin is to be tossed fairly, since a fair coin can be tossed unfairly by some professionals! Experiments can be conducted under different conditions (for example the coin need not be fair), so more than one probability measure can be defined on the same $\mathcal{F}$ and Ω. Often, probability can be interpreted as the frequency with which events would occur if the experiment were to be replicated ad infinitum. To sum up the features of the experiment, the *probability space* or *triplet* $(\Omega, \mathcal{F}, \Pr(\cdot))$ is used.

For any two elements of a sigma-algebra, say A and B, it follows (Exercise 1.5) that

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B).$$

There are two special cases worthy of attention where this formula can be simplified. First, two sets A and B are *mutually exclusive* (for example, $A =$ "raining tomorrow morning" and $B =$ "not raining tomorrow") if and only if the sets are disjoint, in which case $\Pr(A \cap B) = \Pr(\varnothing) = 0$ and hence $\Pr(A \cup B) = \Pr(A) + \Pr(B)$. Second, the sets A and B are *independent* (for example, $A =$ "you catch a cold" and $B =$ "your favorite program will be on TV") if and only if $\Pr(A \cap B) = \Pr(A)\Pr(B)$. If there are three sets A_1, A_2, A_3, we say that they are *pairwise independent* if and only if $\Pr(A_i \cap A_j) = \Pr(A_i)\Pr(A_j)$ for $i = 1, 2$ and $j > i$ (three combinations in all). They are *mutually* (or *jointly*) *independent* if and only if $\Pr(A_1 \cap A_2 \cap A_3) = \Pr(A_1)\Pr(A_2)\Pr(A_3)$. Pairwise independence does not necessarily lead to joint independence, as will be seen in Exercise 1.22. When using the term "independence" in the case of many events, we will mean joint independence unless stated otherwise.

If an event A were to occur, it may convey information about the possibility of realization of another uncertain event B. For example, suppose a teacher is waiting for her students in a lecture theater which has no windows. If several turn up holding wet umbrellas or coats (event A), it is likely that it's been raining outside (event B). The former event has conveyed some information about the latter, even though the latter couldn't be observed directly. The use of information in this way is called *conditioning*: the probability of B

being realized, if A were to occur, is denoted by $\Pr(B \mid A)$. When $\Pr(A) \neq 0$, this *conditional probability* is

$$\Pr(B \mid A) = \frac{\Pr(B \cap A)}{\Pr(A)}, \tag{1.1}$$

as will be seen and generalized in Exercise 1.25. The function $\Pr(B \mid A)$ satisfies the three defining properties of a probability measure on $\mathcal{F}$, which were given earlier.

The formula for conditional probability is important in many ways. First, we can obtain an alternative characterization of the independence of two events A and B as

$$\Pr(B \mid A) \equiv \frac{\Pr(B \cap A)}{\Pr(A)} = \frac{\Pr(B)\Pr(A)}{\Pr(A)} \equiv \Pr(B), \tag{1.2}$$

in which case event A conveys no information about event B, so conditioning on the former is superfluous: $\Pr(B \mid A) = \Pr(B)$. Notice that this definition of independence *seems* to treat A and B in different ways, unlike the earlier definition $\Pr(A \cap B) = \Pr(A)\Pr(B)$ which is symmetric in A and B. However, the same derivations as in (1.2), but with roles reversed, show that $\Pr(A \mid B) = \Pr(A)$ is also the case.

Second, one may apply the conditional factorization twice, when $\Pr(B) \neq 0$ as well, to get

$$\Pr(B \mid A) = \frac{\Pr(B \cap A)}{\Pr(A)} = \frac{\Pr(A \cap B)}{\Pr(A)} = \Pr(A \mid B)\frac{\Pr(B)}{\Pr(A)}, \tag{1.3}$$

which is one form of *Bayes' law*. Before extending this formula, let us introduce the following notation:

$$\bigcap_{i=1}^{n} A_i := A_1 \cap A_2 \cap \cdots \cap A_n$$

and

$$\bigcup_{i=1}^{n} A_i := A_1 \cup A_2 \cup \cdots \cup A_n$$

for the case of a sequence of sets $A_1, \dots, A_n$. If one were to *partition* Ω, that is, to decompose Ω into a collection of some mutually disjoint subsets $C_1, \dots, C_m$ such that

$$\Omega = \bigcup_{i=1}^{m} C_i$$

and $\Pr(C_i) \neq 0$ for all i, then

$$\Pr(B \mid A) = \Pr(A \mid B)\frac{\Pr(B)}{\sum_{i=1}^{m} \Pr(A \mid C_i)\Pr(C_i)}; \tag{1.4}$$

see Exercise 1.6. The sum

$$\sum_{i=1}^{m} \Pr(A \mid C_i)\Pr(C_i) \equiv \sum_{i=1}^{m} \Pr(A \cap C_i) = \Pr(A) \tag{1.5}$$

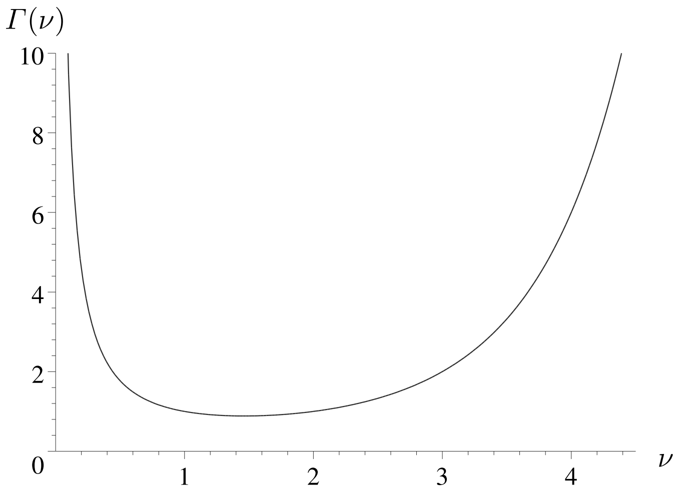

Figure 1.1. Gamma function, $\Gamma(\nu)$.

is sometimes called the *unconditional* (or *marginal*) *probability* of A, to stress the distinction from the conditional probability $\Pr(A \mid C_i)$. As an alternative to (1.3), a second form of Bayes' law occurs when A of (1.4) belongs to a sigma-algebra of events and $B = C_i$ for one specific i, for example, when $B = C_2$. Some illustrations are given in Section 1.2, in particular starting with Exercise 1.26.

Probabilities in this chapter and elsewhere often require computation of the *factorial function*, given by

$$n! = 1 \times 2 \times \cdots \times n$$

when n is a positive integer and $0! = 1$. We summarize this definition by

$$n! := \prod_{i=1}^{n} i, \qquad n = 0, 1, \ldots,$$

where empty products like $\prod_{i=1}^{0}$ are equal to 1, by mathematical convention, so $0! = 1$. This function refers to the number of ways in which n entities can be ordered. For example, there are $3 \times 2 \times 1 = 3!$ ways to order three individuals in a queue: there are three possible choices for assigning one of them to the head of the queue, two possibilities for the next in line, and finally only one remaining individual for the last position. The factorial function satisfies the recursion $n! = n \times (n-1)!$ for $n \in \mathbb{N}$, where $\mathbb{N}$ is the set of natural numbers $1, 2, \ldots$ We now introduce four extensions of this function.

First, the *gamma* (or *generalized factorial*) *function* is

$$\Gamma(\nu) := \int_0^\infty x^{\nu-1} \mathrm{e}^{-x}\, \mathrm{d}x, \qquad \nu > 0, \tag{1.6}$$

and satisfies the recursion $\Gamma(\nu) = (\nu - 1) \times \Gamma(\nu - 1)$ obtained by integrating by parts in (1.6); see Figure 1.1 for its plot. The definition of the function can be extended through this recursion for values of ν that are negative and not integers; however, we shall not need this in the current volume. When ν is a natural number, we obtain the factorial function: $\Gamma(\nu) = (\nu-1)!$. Another important special case of the gamma function is $\Gamma(\frac{1}{2}) = \sqrt{\pi}$, which is represented by the integral

$$\Gamma\left(\frac{1}{2}\right) = \int_0^\infty \mathrm{e}^{-x} \frac{\mathrm{d}x}{\sqrt{x}} = \sqrt{2}\int_0^\infty \mathrm{e}^{-y^2/2}\,\mathrm{d}y$$

by the change of variable $y = \sqrt{2x}$. This result will eventually be proved in Exercise 7.27. It allows the calculation of $\Gamma(n+\frac{1}{2})$ by recursion for all $n \in \mathbb{N}$.

Second, the *beta function* is defined by

$$B(\nu,\mu) := \frac{\Gamma(\nu)\Gamma(\mu)}{\Gamma(\nu+\mu)} = B(\mu,\nu),$$

and, when $\nu, \mu > 0$, we have two equivalent integral representations of the function,

$$B(\nu,\mu) = \int_0^1 x^{\nu-1}(1-x)^{\mu-1}\,\mathrm{d}x = \int_0^\infty \frac{y^{\nu-1}}{(1+y)^{\nu+\mu}}\,\mathrm{d}y,$$

by setting $x = y/(1+y)$.

The final two extensions are very closely linked and have an important probabilistic interpretation. For $j = 0, 1, \dots$, the j (ordered) *permutations* of ν are

$$\mathrm{P}_j^\nu := \prod_{i=0}^{j-1}(\nu - i) = (\nu)(\nu-1)\cdots(\nu-j+1) = \frac{\Gamma(\nu+1)}{\Gamma(\nu-j+1)},$$

and the j (unordered) *combinations* of ν are

$$\binom{\nu}{j} := \frac{\prod_{i=0}^{j-1}(\nu-i)}{j!} = \frac{\Gamma(\nu+1)}{\Gamma(\nu-j+1)j!},$$

where $\binom{\cdot}{\cdot}$ is the symbol for the *binomial coefficient*, sometimes written as C_j^ν. These two functions are generally defined for $\nu \in \mathbb{R}$, which we will require later but not in this chapter. Here, we deal with the special case $\nu = n \in \mathbb{N}$ yielding

$$\frac{\Gamma(n+1)}{\Gamma(n-j+1)} = \frac{n!}{(n-j)!}.$$

In this case, the definition of binomial coefficients implies directly that

$$\binom{n}{j} = \frac{n!}{(n-j)!j!} = \binom{n}{n-j}. \tag{1.7}$$

Continuing with our earlier example of individuals in queues, suppose that we want to form a queue of $j = 3$ individuals from a group of n people, where $n \geqslant 3$. Then, we can do this in $n \times (n-1) \times (n-2) = \mathrm{P}_3^n$ ways. Now, suppose instead that we select the three individuals simultaneously and that ordering does not matter, for example because all three customers can be served simultaneously. Since there are 3! ways to rearrange any selection of three specific individuals, the number of ways to select three simultaneously is $\mathrm{P}_3^n/3! = \binom{n}{3}$. It is also equal to the number of ways to select $n-3$ individuals (or leave out three of them), $\binom{n}{n-3}$, a result implied more generally by (1.7). Generalizing these ideas to selecting $j_1, \dots, j_k$ individuals from respective groups of $n_1, \dots, n_k$ people, there are

$$\binom{n_1}{j_1} \cdots \binom{n_k}{j_k}$$

unordered selections; for example, selecting one from the set of two individuals $\{R, J\}$ and one from $\{K\}$, we can have $\binom{2}{1}\binom{1}{1} = 2$ (unordered) combinations: R, K or J, K. For ordered selections, we have as many as

$$\binom{n_1}{j_1} \cdots \binom{n_k}{j_k} \times (j_1 + \cdots + j_k)! = \mathrm{P}^{n_1}_{j_1} \cdots \mathrm{P}^{n_k}_{j_k} \times \frac{(j_1 + \cdots + j_k)!}{j_1! \cdots j_k!},$$

where $(j_1 + \cdots + j_k)!$ is the number of ways to order $j_1 + \cdots + j_k$ individuals. In the latter equation, we can interpret $\mathrm{P}^{n_\bullet}_{j_\bullet}$ as the number of ordered selections *within* each of the k groups, whereas

$$\frac{(j_1 + \cdots + j_k)!}{j_1! \cdots j_k!} \tag{1.8}$$

is the number of ways of allocating slots (say, in a queue of $j_1 + \cdots + j_k$) to groups, without distinction of the individuals within each group (by selecting j_1 simultaneously from group 1, and so on). Continuing with our last example, one from $\{\dagger, \dagger\}$ and one from $\{\star\}$ can be arranged in $(1 + 1)!/(1!1!) = 2$ ways, as $\dagger, \star$ or $\star, \dagger$. The factor in (1.8) is called the *multinomial coefficient* because it generalizes the binomial coefficient obtained when $k = 2$, which makes it particularly useful from Chapter 5 onwards.

The exercises in this chapter follow broadly the sequence of topics introduced earlier. We start with illustrations of random experiments and probabilities, then move on to conditioning. We conclude with a few exercises focusing on permutations and combinations.

1.1 Events and sets

Exercise 1.1 (Urn) An urn contains m red, m white, and m green balls ($m \geqslant 2$). Two balls are drawn at random, without replacement.
(a) What is the sample space?
(b) Define the events $A :=$ "drawing a green ball first" and $B :=$ "drawing at least one green ball". Express A and B as unions of elementary events.
(c) Also express $A \cap B$ and $A^c \cap B$ as unions of elementary events.

Solution
(a) Denote the red, white, and green balls by R, W, and G, respectively. Since the order matters, the sample space contains nine elements: $\Omega = \{RR, RW, RG, WR, WW, WG, GR, GW, GG\}$.
(b) $A = \{GR, GW, GG\}$ and $B = \{RG, WG, GR, GW, GG\}$.

(c) $A \cap B$ contains the elements that are in both A and B. Hence, $A \cap B = \{GR, GW, GG\}$. Notice that $A \cap B = A$ since A is a subset of B. $A^c \cap B$ contains the elements that are in B but not in A: $A^c \cap B = \{RG, WG\}$.

Exercise 1.2 (Urn, continued) Consider again the experiment of Exercise 1.1.
(a) Do the elements in the sample space have equal probability?
(b) Compute $\Pr(A)$ and $\Pr(B)$.
(c) Are A and B independent?

Solution
(a) If the two balls had been drawn with replacement, the sample space elements would have had equal probability $1/9$ each. However, without replacement, we have, for $i, j = R, W, G$,

$$\Pr(ii) = \frac{m}{3m} \times \frac{m-1}{3m-1} = \frac{m-1}{9m-3} < \frac{1}{9}$$

and, for $i \neq j$,

$$\Pr(ij) = \frac{m}{3m} \times \frac{m}{3m-1} = \frac{m}{9m-3} > \frac{1}{9}.$$

We see that both $\Pr(ii)$ and $\Pr(ij)$ approach $1/9$ when $m \to \infty$. Notice that we have adopted a shorthand notation that drops the braces around ij when it appears inside $\Pr(\cdot)$, a simplification used from now on.
(b) We have

$$\Pr(A) = \Pr(GR) + \Pr(GW) + \Pr(GG) = \frac{2m}{9m-3} + \frac{m-1}{9m-3} = \frac{1}{3}$$

and

$$\Pr(B) = \frac{4m}{9m-3} + \frac{m-1}{9m-3} = \frac{5m-1}{9m-3} > \frac{5}{9}.$$

The first result is also immediately obtained from $\Pr(A) = m/(3m)$.
(c) The two events are certainly not independent because A is a subset of B. As a result, $\Pr(A \cap B) = \Pr(A) > \Pr(A)\Pr(B)$.

Exercise 1.3 (Coin) Hedda tosses a fair coin four times.
(a) Give the sample space.
(b) What is the probability that she throws exactly three heads?
(c) What is the probability that she throws at least one head?
(d) What is the probability that the number of heads exceeds the number of tails?
(e) What is the probability that the number of heads equals the number of tails?

Solution
(a) Since each toss has two possible outcomes, there are $2^4 = 16$ sample elements and the sample space is $\Omega = \{HHHH, HHHT, HHTH, HTHH, THHH, TTHH, THTH,$