PART I

PROBABILITY

CHAPTER 1

ELEMENTS OF PROBABILITY MEASURE

The axiomatic approach of Kolmogorov is followed by most books on probability theory. This is the approach of choice for most graduate level probability courses. However, the immediate applicability of the theory learned as such is questionable, and many years of study is required to understand and unleash its full power.

On the other hand, the books on applied probability completely disregard this approach, and they go more or less directly into presenting applications, thus leaving gaps in the reader's knowledge. On a cursory glance, this approach appears to be very useful (the presented problems are all very real and most are difficult). However, I question the utility of this approach when confronted with problems that are slightly different from the ones presented in such books.

I believe no present textbook strikes the right balance between these two approaches. This book is an attempt in this direction. I will start with the axiomatic approach and present as much as I feel will be be necessary for a complete understanding of the theory of probability. I will skip proofs which I consider will not bring something new to the development of the student's understanding.

Probability and Stochastic Processes, First Edition. Ionuţ Florescu

1.1 Probability Spaces

Let Ω be an abstract set containing all possible outcomes or results of a random experiment or phenomenon. This space is sometimes denoted with S and is named the *sample space*. I call it "an abstract set" because it could contain anything. For example, if the experiment consists in tossing a coin once, the space Ω could be represented as $\{Head, Tail\}$. However, it could just as well be represented as $\{Cap, Pajura\}$, these being the Romanian equivalents of $Head$ and $Tail$. The space Ω could just as well contain an infinite number of elements. For example, measuring the diameter of a doughnut could result in all possible numbers inside a whole range. Furthermore, measuring in inches or in centimeters would produce different albeit equivalent spaces.

We will use ω, where $\omega \in \Omega$ to denote a generic outcome or a sample point.

Any collection of outcomes is called an event. That is, any subset of Ω is an event. We shall use capital letters from the beginning of the alphabet ($A, B, C, \ldots$) to denote events.

So far so good. The proper definition of Ω is one of the most important issues when treating a problem probabilistically. However, this is not enough. We have to make sure that we can calculate the probability of all the items of interest.

Think of the following possible situation: Poles of various sizes are painted in all possible colors. In other words, the poles have two characteristics of interest: size and color. Suppose that in this model we have to calculate the probability that the next pole would be shorter than 15 in. and painted either red or blue. In order to answer such questions, we have to properly define the sample space Ω and, furthermore, give a definition of probability such that the calculations are consistent. Specifically, we need to describe the elements of Ω which **can be** measured.

To this end, we have to group these events in some way that would allow us to say: yes, we can calculate the probability of all the events in this group. In other words, we need to talk about the notion of a collection of events.

We will introduce the notion of σ-algebra (or σ-field) to deal with the problem of the proper domain of definition for the probability. Before we do that, we introduce a special collection of events:

$$\mathscr{P}(\Omega) = \text{The collection of all possible subsets of } \Omega \tag{1.1}$$

We could define probability on this very large set. However, this would mean that we would have to define probability for every single element of $\mathscr{P}(\Omega)$. This will prove impossible except in the case when Ω is finite. In this case, the collection $\mathscr{P}(\Omega)$ is called *the power set of* Ω. However, even in this case we have to do it consistently. For example, if, say, the set $\{1, 2, 3\}$ is in Ω and has probability 0.2, how do we define the probability of $\{1, 2\}$? How about the probability of $\{1, 2, 5\}$? A much better approach would be to define probability for a smaller set of important elements which generate the collection $\mathscr{P}(\Omega)$. For example, if we define the probability of each of the generators $1, 2, \ldots, 5$, perhaps we can then say something about the probabilities of the bigger events.

How do we do this? Fortunately, algebra comes to our rescue. The elements of a collection of events are the events. So, first we define operations with them: *union, intersection, complement* and slightly less important *difference and symmetric difference*.

$$\begin{cases} A \cup B & = \text{set of elements that are } \textbf{either} \text{ in } A \textbf{ or} \text{ in } B \\ A \cap B & = AB = \text{set of elements that are } \textbf{both} \text{ in } A \textbf{ and} \text{ in } B \\ A^c & = \bar{A} = \text{set of elements that are in } \Omega \text{ but } \textbf{not} \text{ in } A \end{cases} \tag{1.2}$$

$$\begin{cases} A \setminus B = & \text{set of elements that are in } A \text{ but } \textbf{not} \text{ in } B \\ A \triangle B = & (A \setminus B) \cup (B \setminus A) \end{cases}$$

We will also use the notation $A \subseteq B$ to denote the case when all the elements in A are also in B, and we say A is a subset of B. $A \subset B$ will denote the case when A is a proper subset of B, that is, B contains at least one other element besides those in A.

We can express every such operation in terms of union and intersection, or basically reduce to only two operations. For example, $A \setminus B = A \cap B^c$. Union may be expressed in terms of intersection, and vice versa. In fact, since these relations are very important, let us state them separately.

De Morgan laws

$$\begin{cases} (A \cup B)^c & = A^c \cap B^c \\ (A \cap B)^c & = A^c \cup B^c \end{cases} \tag{1.3}$$

We will mention one more property, which is the distributivity property, that is, union and intersection distribute to each other:

Distributivity property of union/intersection

$$\begin{aligned} A \cup (B \cap C) &= (A \cup B) \cap (A \cup C) \\ A \cap (B \cup C) &= (A \cap B) \cup (A \cap C) \end{aligned} \tag{1.4}$$

There is much more to be found about set operations, but for our purpose this is enough. We recommend you to look up Billingsley (1995) or Chung (2000) for a wealth of details.

Definition 1.1 (Algebra on Ω) *A collection $\mathscr{F}$ of events in Ω is called an algebra (or field) on Ω iff*

1. $\Omega \in \mathscr{F}$;
2. *Closed under complementarity: If $A \in \mathscr{F}$ then $A^c \in \mathscr{F}$;*
3. *Closed under finite union: If $A, B \in \mathscr{F}$ then $A \cup B \in \mathscr{F}$.*

Remark 1.2 *The first two properties imply that* $\varnothing \in \mathscr{F}$. *The third is equivalent to* $A \cap B \in \mathscr{F}$ *by the second property and the de Morgan laws* (1.3).

Definition 1.3 (σ-Algebra on Ω) *If* $\mathscr{F}$ *is an algebra on* Ω *and, in addition, it is closed under countable unions, then it is a* σ*-algebra (or* σ*-field) on* Ω.

Note Closed under countable unions means that the third property in Definition 1.1 is replaced with the following: If $n \in \mathbb{N}$ is a natural number and $A_n \in \mathscr{F}$ for all n, then

$$\bigcup_{n\in\mathbb{N}} A_n \in \mathscr{F}.$$

The σ-algebra provides an appropriate domain of definition for the probability function. However, it is such an abstract thing that it will be hard to work with it. This is the reason for the next definition. It will be much easier to work with the generators of a $sigma$-algebra. *This will be a recurring theme in probability. In order to show a property for a certain category of objects, we show that the property holds for a small set of objects which generate the entire class. This will be enough – standard arguments will allow us to extend the property to the entire category of objects.*

EXAMPLE 1.1 A simple example of σ-algebra

Suppose a set $A \subset \Omega$. Let us calculate $\sigma(A)$. Clearly, by definition Ω is in $\sigma(A)$. Using the complementarity property, we clearly see that A^c and $\emptyset$ are also in $\sigma(A)$. We only need to take unions of these sets and see that there are no more new sets. Thus

$$\sigma(A) = \{\Omega, \emptyset, A, A^c\}.$$

■

Definition 1.4 (σ algebra generated by a class $\mathscr{C}$ of sets in Ω) *Let* $\mathscr{C}$ *be a collection (class) of sets in* Ω. *Then* $\sigma(\mathscr{C})$ *is the smallest* σ*-algebra on* Ω *that contains* $\mathscr{C}$.

Mathematically

1. $\mathscr{C} \subseteq \sigma(\mathscr{C})$,
2. $\sigma(\mathscr{C})$ *is a* σ*-field,*
3. *If* $\mathscr{C} \subseteq \mathscr{G}$ *and* $\mathscr{G}$ *is a* σ*-field, then* $\sigma(\mathscr{C}) \subseteq \mathscr{G}$.

As we have mentioned, introducing generators of the large collection of sets $\sigma(\mathscr{C})$ is fundamental. Suppose that we want to show that a particular statement is true for every set in the large collection. If we know the generating subset $\mathscr{C}$, we can verify whether the statement is true for the elements of $\mathscr{C}$, which is a much easier task. Then,

because of the properties that would be presented later, the particular statement will be valid for all the sets in the larger collection $\sigma(\mathscr{C})$.

Proposition 1.5 *Properties of σ-algebras:*

- $\mathscr{P}(\Omega)$ *is the largest possible σ-algebra defined on* Ω*;*
- *If* $\mathscr{C}$ *is already a σ-algebra, then* $\sigma(\mathscr{C}) = \mathscr{C}$*;*
- *If* $\mathscr{C} = \{\varnothing\}$ *or* $\mathscr{C} = \{\Omega\}$*, then* $\sigma(\mathscr{C}) = \{\varnothing, \Omega\}$*, the smallest possible σ-algebra on* Ω*;*
- *If* $\mathscr{C} \subseteq \mathscr{C}'$*, then* $\sigma(\mathscr{C}) \subseteq \sigma(\mathscr{C}')$*;*
- *If* $\mathscr{C} \subseteq \mathscr{C}' \subseteq \sigma(\mathscr{C})$*, then* $\sigma(\mathscr{C}') = \sigma(\mathscr{C})$.

EXAMPLE 1.2

Suppose that $\mathscr{C} = \{A, B\}$, where A and B are two sets in Ω such that $A \subset B$. Let us list the sets in $\sigma(\mathscr{C})$. A common mistake made by students is the following argument:

$A \subset B$, therefore using the fourth property in the proposition above $\sigma(A) \subseteq \sigma(B)$ and therefore the sigma algebra asked is $\sigma(A, B) = \{\Omega, \emptyset, B, B^c\}$, done.

This argument is wrong on several levels. First, the quoted property refers to collections of sets and not to the sets themselves. While it is true that $A \subset B$, it is not true that $\{A\} \subset \{B\}$ as collections of sets. Instead, $\{A\} \subset \{A, B\}$ and, indeed, this implies $\sigma(A) \subseteq \sigma(A, B)$. But this just means that the result should contain all the sets in $\sigma(A)$ (the sets in the previous example). Furthermore, as the example will show and as the Proposition 1.6 says, it isn't true either that $\sigma(A) \cup \sigma(B) = \sigma(A, B)$. The only way to solve this problem is the hard way.

Clearly, $\sigma(A, B)$ should contain the basic sets and their complements, thus $\sigma(A, B) \supset \{\Omega, \emptyset, A, B, A^c, B^c\}$. It should also contain all their unions according to the definition. Therefore, it must contain

$$A \cup B = B$$
$$A \cup B^c$$
$$A^c \cup B = \Omega$$
$$A^c \cup B^c = A^c$$

where the equalities are obtained using that $A \subset B$. So the only new set to be added is $A \cup B^c$ and thus its complement as well: $(A \cup B^c)^c = A^c \cap B$. Now we need to verify that by taking unions we do not obtain any new sets, a task left for the reader. In conclusion, when $A \subset B$,

$$\sigma(A, B) = \{\Omega, \emptyset, A, B, A^c, B^c, A \cup B^c, A^c \cap B\}$$

Proposition 1.6 (Intersection and union of σ-algebras) *Suppose that $\mathscr{F}_1$ and $\mathscr{F}_2$ are two σ-algebras on Ω. Then*

1. *$\mathscr{F}_1 \cap \mathscr{F}_2$ is a sigma algebra.*

2. *$\mathscr{F}_1 \cup \mathscr{F}_2$ is **not** a sigma algebra. The smallest σ algebra that contains both of them is $\sigma(\mathscr{F}_1 \cup \mathscr{F}_2)$ and is denoted $\mathscr{F}_1 \vee \mathscr{F}_2$.*

Proof: For part 2, there is nothing to show. A counterexample is provided by the example 1.2. Take $\mathscr{F}_1 = \sigma(A)$ and $\mathscr{F}_2 = \sigma(B)$. The example above calculates $\sigma(A, B)$ and it is simple to see that $\mathscr{F}_1 \cup \mathscr{F}_2$ needs more sets to become a sigma algebra (e.g., $A \cup B^c$).

For part 1, we just need to verify the definition of the sigma algebra. Take a set A in $\mathscr{F}_1 \cap \mathscr{F}_2$. So A belongs to both collections of sets. Since $\mathscr{F}_1$ is a sigma algebra by definition, $A^c \in \mathscr{F}_1$. Similarly, $A^c \in \mathscr{F}_2$. Therefore, $A^c \in \mathscr{F}_1 \cap \mathscr{F}_2$. The rest of the definition is verified in a similar manner. ∎

Finite, countable, uncountable (infinite) sets

In general, listing the elements of a sigma algebra explicitly is hard. It is only in simple cases that this can be done. Specifically, we can do this in the case of finite and countable spaces Ω. But what exactly is a countable space and how does it differ from uncountable spaces? Let us present some simple definitions to this end.

Definition 1.7 (Injective, surjective, bijective functions) *Let $f : \Omega \longrightarrow \Gamma$ be some function. We say that the function f is*

*i) **Injective** or **one to one** if $\forall x \neq y$ both in Ω we have $f(x) \neq f(y)$. Injectivity is usually checked by the equivalent statement "$f(x) = f(y) \Rightarrow x = y$".*

*ii) **Surjective** or **onto** if for all $y \in \Gamma$, there exist at least an $x \in \Omega$ such that $f(x) = y$. This is normally checked by picking an arbitrary y in Γ and actually finding the x that goes into it. This procedure gives the set $f^{-1}(\{y\})$.*

*iii) **Bijective** if it is both one to one and onto. The function is a bijection if the inverse set $f^{-1}(\{y\})$ has exactly one element for all y. In this case, f^{-1} is a proper function defined on Γ with values in Ω called the* inverse function.

Definition 1.8 (Cardinality of a set) *Suppose that we can construct a bijection between two sets Ω and Γ. Then we say that the cardinality of the two sets is the same (denoted $|\Omega| = |\Gamma|$).*

Important sets:

$$\mathbb{N} = \{0, 1, 2, 3, \ldots\} = \text{natural numbers}$$
$$\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\} = \text{integer numbers}$$
$$\mathbb{Q} = \left\{ \frac{m}{n} \mid m, n \in \mathbb{Z} \right\} = \text{rational numbers}$$
$$\mathbb{R} = \text{real numbers } (\sigma\text{-field on } \mathbb{Q} \text{ under addition and multiplication})$$
$$\mathbb{C} = \{a + bi \mid a, b \in \mathbb{R}\} = \text{complex numbers}$$

We shall denote with a star any set not containing 0. For example, $\mathbb{Z}^* = \mathbb{Z} \backslash \{0\}$.

Suppose $n \in \mathbb{N}$ is some natural number. We define the cardinality of the set $A = \{1, 2, \ldots, n\}$ as $|A| = n$. We define the cardinality of $\mathbb{N}$ with $\aleph_0$. I will not get into details about these *aleph* numbers. For details, refer to an introductory textbook such as Halmos (1998).

Definition 1.9 *Any set with cardinality equal to a finite number $n \in \mathbb{N}$ is called a* finite set. *Any set with cardinality $\aleph_0$ is called a* countable set. *Any set which is not one of the first two is called* uncountable *(infinite).*

The definition of infinite is a bit strange ($\mathbb{N}$ has an infinite number of elements), but it is important to us because it helps us to make the distinction between say $\mathbb{Q}$ and any interval, say $(0.00000000000000001, 0.00000000000000002)$. This tiny interval has a lot more elements than the entire $\mathbb{Q}$. I will give two more results then we return to probability.

Proposition 1.10 *The following results hold:*

1) $|\mathbb{N}| = |\mathbb{Z}|$

2) $|\mathbb{N}| = |\mathbb{Q}|$

3) $|\mathbb{R}| = |(0, 1)|$

4) $\mathbb{R}$ *is uncountable (infinite).*

Proof: The proof is very simple and amusing, so I am giving it here.

1) Let $f : \mathbb{N} \to \mathbb{Z}$, which takes odd numbers in negative integers and even numbers in positive integers, that is

$$f(x) = \begin{cases} x/2, & \text{if } x \text{ is even} \\ -(x+1)/2, & \text{if } x \text{ is odd} \end{cases}$$

It is easy to show that this function is a bijection between $\mathbb{N}$ and $\mathbb{Z}$; thus, according to the definition, the two have the same cardinality ($\mathbb{Z}$ is countable).

2) From definition

$$\mathbb{Q} = \bigcup \left\{ \frac{m}{n} \middle| m, n \in \mathbb{Z} \right\} = \bigcup_{n \in \mathbb{Z}} \left\{ \frac{m}{n} \middle| m \in \mathbb{Z} \right\} = \bigcup_{n \in \mathbb{Z}} Q_n,$$

where we use the notation $Q_n = \left\{ \frac{m}{n} \middle| \, m \in \mathbb{Z} \right\}$. For any fixed n, there clearly exists a bijection between $\mathbb{Z}$ and Q_n ($g(x) = x/n$). Thus Q_n is countable. But then $\mathbb{Q}$ is a countable union of countable sets. Lemma 1.11 proved next shows that the result is always a countable set (the proof is very similar to the proof above of $\mathbb{Z} = \mathbb{N} \cup (-\mathbb{N})$).

3) Let $f : \mathbb{R} \to (0, 1)$,
$$f(x) = \frac{e^x}{1 + e^x}.$$
This function is a bijection (exercise).

4) From the previous point 3, it is enough to show that $(0, 1)$ is uncountable. Assume by absurd that the interval is countable. Then there must exist a bijection with $\mathbb{N}$ and therefore its elements may be written as a sequence:
$$\begin{aligned} x_1 &= 0.x_{11}x_{12}x_{13}\ldots \\ x_2 &= 0.x_{21}x_{22}x_{23}\ldots \\ x_3 &= 0.x_{31}x_{32}x_{33}\ldots \\ &\vdots \quad \vdots \end{aligned}$$
where x_{ij} are the digits of the number x_i. To finish the reduction to absurd, we construct a number which is not on the list. This number is going to be different than each of the numbers listed in at least one digit. Specifically, construct the number $y = 0.y_{11}y_{22}y_{33}\ldots$, such that the digit i is
$$y_{ii} = \begin{cases} 2, & \text{if } x_{ii} = 1 \\ 1, & \text{if } x_{ii} \neq 1 \end{cases}$$
The point is that the digit i of the number y is different from the digit i of the number x_i, and therefore the two numbers cannot be equal. This is happening for all the i's, so y is a new number not on the list. This contradicts the assumption that we may list all the numbers in the interval. Thus $(0, 1)$ is not countable.

■

In the above proof, we need the following lemma which is very useful by itself.

Lemma 1.11 *Suppose $A_1, A_2, \ldots$ are countable sets. Then $\cup_{n\geq 1} A_n$ is a countable set.*

Proof: The proof basically uses the fact that a composition of two bijections is a bijection. Suppose we look at one of these sets say A_n. Since this set is countable, we may list its elements as $A_n = \{a_{m1}, a_{m2}, a_{m3}, \ldots\}$, and similarly for all the sets. Now consider the function $f : \cup_{n\geq 1} A_n \to \mathbb{N}$, with
$$f(a_{mn}) = \frac{(m + n - 1)(m + n - 2)}{2} + m$$

This function basically assigns a_{11} to 1, a_{12} to 2, a_{21} to 3, a_{22} to 5, and so on.

The function f is clearly an injection because no two elements are taken at the same point, but it may not be surjective (for example, some set A_i may be finite and thus those missing elements a_{ij} may not be taken into integers $\frac{(i+j-1)(i+j-2)}{2} + i$ and thus these integers have $f^{-1}(\{y\}) = \emptyset$). However, we can restrict the codomain to the image of the function, and thus the set $\cup_{n\geq 1} A_n$ has the same cardinality as a set included in $\mathbb{N}$ but any subset of $\mathbb{N}$ is countable (easy to show), done. ∎

Finally, the following result is relevant for the next chapter. The lemma may be applied to the distribution function (a nondecreasing bounded function).

Lemma 1.12 *Let $f : \mathbb{R} \to \mathbb{R}$ be a nondecreasing function. Then the set of points where f has a discontinuity is countable.*

Proof: Let A the set of discontinuity points of f. We need to show that A is countable. Let $x \in A$. The function is discontinuous at x, so the left and right limits are different: $f(x-) = \lim_{y\uparrow x} f(y) \neq \lim_{y\downarrow x} f(y) = f(x+)$. Since f is increasing and the two values are different, there must exist some number $q_x \in \mathbb{Q}$ such that

$$f(x-) < q_x < f(x+).$$

This allows us to construct a function $g : A \to \mathbb{Q}$, by $g(x) = q_x$. This function is injective. Why? Let $x \neq y \in A$. Then we have, say, $x < y$. Since f is increasing, we have $f(x) < f(y)$. But then we must also have

$$q_x < f(x+) < f(y-) < q_y$$

Thus $q_x \neq q_y$, and so the function g is injective. Again, like we did in the previous lemma, we restrict the codomain to the image of A, which makes the function a bijection. Therefore, the set A is countable. ∎

Back to σ-algebras.

Remark 1.13 (Very important. The case of countable space Ω) *When the sample space is finite, we can and typically will take the sigma algebra to be $\mathscr{P}(\Omega)$. Indeed, any event of a finite space can be trivially expressed in terms of individual outcomes. In fact, if the space Ω is finite, say it contains m possible outcomes, then the number of possible events is finite and is equal to 2^m.*

Remark 1.13 is important. It basically says that if you are looking to understand the concept and usefulness of sigma algebras, looking at coin tosses or rolling of dies is irrelevant. For these simple experiments, the notion of σ-algebra is useless because we can just take the events that can be measured as all the possible events. σ-algebras are useful even in these simple examples if we start thinking about tossing the coin or rolling the die ad infinitum. That sample space is not finite. We next present the Borel σ-algebra, which is going to be essential for understanding random variables.

An example: Borel σ-algebra

Let Ω be a topological space (think geometry is defined in this space and this assures us that the open subsets exist in this space)[1].

Definition 1.14 *We define*

$$\begin{aligned}\mathscr{B}(\Omega) &= \textit{The Borel } \sigma\textit{-algebra} \\ &= \sigma\textit{-algebra generated by the class of open subsets of } \Omega\end{aligned} \tag{1.5}$$

In the special case when $\Omega = \mathbb{R}$, we denote $\mathscr{B} = \mathscr{B}(\mathbb{R})$, the Borel sets of $\mathbb{R}$. This $\mathscr{B}$ is the most important σ-algebra. The reason for this is that most experiments can be brought to equivalence with $\mathbb{R}$ (as we shall see when we will talk about random variables). Thus, if we define a probability measure on $\mathscr{B}$, we have a way to calculate probabilities for most experiments. ■

Most subsets of $\mathbb{R}$ are in $\mathscr{B}$. However, it is possible (though very difficult) to explicitly construct a subset of $\mathbb{R}$ which is not in $\mathscr{B}$. See (Billingsley, 1995, page 45) for such a construction in the case $\Omega = (0, 1]$.

There is nothing special about the open sets, except for the fact that they can be defined in any topological space (thus they always exist given a topology). In $\mathbb{R}$, we can generate the Borel σ-algebra using many classes of generators. In the end, the same σ-algebra is reached (see problem 1.8).

Probability measure

We are finally in a position to give the domain for the probability measure.

Definition 1.15 (Measurable space) *A pair $(\Omega, \mathscr{F})$, where Ω is a set and $\mathscr{F}$ is a σ-algebra on Ω, is called a* measurable space.

Definition 1.16 (Probability measure. Probability space) *Given a measurable space $(\Omega, \mathscr{F})$, a probability measure is any function $\mathbf{P} : \mathscr{F} \to [0, 1]$ with the following properties:*

i) $\mathbf{P}(\Omega) = 1$

[1] For completion, we present the definition of a topological space even though it is secondary to our purpose. A topological space is a set Ω together with a collection of subsets of Ω, called open sets, and satisfying the following three axioms:

1. The empty set and Ω itself are open;
2. Any union of open sets is open;
3. The intersection of any finite number of open sets is open.

This collection of open sets (denoted $\mathscr{T}$) is called a *topology on* Ω or an *open set topology*. The sets in $\mathscr{T}$ are called *open sets*, and their complements in Ω are called *closed sets*. A subset of Ω may be neither closed nor open, either closed or open, or both.

ii) (countable additivity) For any sequence $\{A_n\}_{n\in\mathbb{N}}$ *of disjoint events in* $\mathscr{F}$ *(i.e.,* $A_i \cap A_j = \varnothing$*, for all* $i \neq j$*):*

$$\mathbf{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbf{P}(A_n).$$

The triple $(\Omega, \mathscr{F}, \mathbf{P})$ *is called a* probability space.

Note that the probability measure is a set function (i.e., a function defined on sets).

The next two definitions define slightly more general measures than the probability measure. We will use these notions later in this book in the hypotheses of some theorems to show that the results apply to more general measures.

Definition 1.17 (Finite measure) *Given a measurable space* $(\Omega, \mathscr{F})$*, a finite measure is a set function* $\mu : \mathscr{F} \to [0, 1]$ *with the same countable additivity property as in Definition 1.16, but the measure of the space is a finite number (not necessarily* 1*). Mathematically, the first property in Definition 1.16 is replaced with*

$$\mu(\Omega) < \infty$$

Note that we can always construct a probability measure from a finite measure μ. For any set $A \in \mathscr{F}$, define a new measure by

$$\mathbf{P}(A) = \frac{\mu(A)}{\mu(\Omega)}.$$

So this notion isn't more general than the probability measure. However, the next notion is.

Definition 1.18 (σ-finite measure) *A measure* μ *defined on a measurable space* $(\Omega, \mathscr{F})$ *is called* σ*-finite if it is countably additive and there exists a partition of the space* Ω *with sets* $\{\Omega_i\}_{i\in I}$*, and* $\mu(\Omega_i) < \infty$ *for all* $i \in I$*. Note that the index set* I *is allowed to be countable.*

To be precise, a partition of any set Ω is any collection of sets $\{\Omega_i\}_{i\in I}$, which are disjoint (i.e., $\Omega_i \cap \Omega_j = \emptyset$, if $i \neq j$) such that their union recreates the original set: $\cup_{i\in I}\Omega_i = \Omega$. The partition may be finite or infinite depending on whether the number of sets in the partition is finite or countable.

As an example of the difference between finite and sigma finite measure spaces, consider the interval $[0, 1)$ and assume that a probability measure is defined on that space (we will talk later about the Lebesgue measure generated by the length of intervals). Typically, such a measure may be extended to any interval of length 1: $[a, a+1)$. But note that the measure will become infinite when extended to $\mathbb{R}$. However, since we can write $\mathbb{R} = \cup_{n\in\mathbb{Z}}[n, n + 1)$, we can see that the measure can be extended to a sigma-finite measure on $\mathbb{R}$. Typically, sigma-finite measures have the same nice properties of probability measures.

EXAMPLE 1.3 Discrete probability space

Let Ω be a countable space. Let $\mathscr{F} = \mathscr{P}(\Omega)$. Let $p : \Omega \to [0, N]$ be a function on Ω such that $\sum_{\omega\in\Omega} p(\omega) = N < \infty$, where N is a finite constant. Define a probability measure by

$$\mathbf{P}(A) = \frac{1}{N}\sum_{\omega\in A} p(\omega) \tag{1.6}$$

We can show that $(\Omega, \mathscr{F}, \mathbf{P})$ is a probability space. Indeed, from the definition (1.6),

$$\mathbf{P}(\Omega) = \frac{1}{N}\sum_{\omega\in\Omega} p(\omega) = \frac{1}{N}N = 1.$$

To show the countable additivity property, let A be a set in Ω such that $A = \bigcup_{i=1}^{\infty} A_i$, with A_i disjoint sets in Ω. Since the space is countable, we may write $A_i = \{\omega_1^i, \omega_2^i, \ldots\}$, where any of the sets may be finite, but $\omega_j^i \neq \omega_l^k$ for all i, j, k, l where either $i \neq k$ or $j \neq l$. Then using the definition (1.6) we have

$$\begin{aligned}\mathbf{P}(A) &= \frac{1}{N}\sum_{\omega\in\bigcup_{i=1}^{\infty} A_i} p(\omega) = \frac{1}{N}\sum_{i\geq 1, j\geq 1} p(\omega_j^i)\\ &= \frac{1}{N}\sum_{i\geq 1}\left(p(\omega_1^i) + p(\omega_2^i) + \ldots\right) = \sum_{i\geq 1}\mathbf{P}(A_i)\end{aligned}$$

■

This is a very simple example, but it shows the basic probability reasoning.

Remark 1.19 *Example 1.3 gives a way to construct discrete probability measures (distributions). For example, take $\Omega = \mathbb{N}$ (the natural numbers) and take $N = 1$ in the definition of the probability of an event. Then using various probability assignments, $p(\omega)$ produces distributions with which you may already be familiar.*

- $p(\omega) = \begin{cases} 1-p & \text{, if } \omega = 0 \\ p & \text{, if } \omega = 1 \\ 0 & \text{, otherwise} \end{cases}$ *, gives the Bernoulli(p) distribution.*
- $p(\omega) = \begin{cases} \binom{n}{\omega} p^{\omega}(1-p)^{n-\omega} & \text{, if } \omega \leq n \\ 0 & \text{, otherwise} \end{cases}$ *, gives the binomial(n,p) distribution.*
- $p(\omega) = \begin{cases} \binom{\omega-1}{r-1} p^{r}(1-p)^{\omega-r} & \text{, if } \omega \geq r \\ 0 & \text{, otherwise} \end{cases}$ *, gives the negative binomial(r,p) distribution.*
- $p(\omega) = \frac{\lambda^{\omega}}{\omega!}e^{-\lambda}$*, gives the Poisson ($\lambda$) distribution.*

EXAMPLE 1.4 Uniform distribution on (0,1)

Let $\Omega = (0,1)$ and $\mathscr{F} = \mathscr{B}((0,1))$ be the Borel sigma algebra generated by the open sets. Define a probability measure U as follows: for any open interval $(a,b) \subseteq (0,1)$, let $U((a,b)) = b - a$ the length of the interval. For any other open interval O, define $U(O) = U(O \cap (0,1))$.

Note that we did not specify the measure $U(A)$ for all of the Borel sets $A \in \mathscr{B}$, but rather only for the generators of the Borel σ-field. This illustrates the probabilistic concept discussed after Definition 1.4 where we define or prove a statement using only the generators of a sigma algebra.

In our specific situation, under very mild conditions on the generators of the σ-algebra, any probability measure defined only on the generators can be uniquely extended to a probability measure on the whole σ-algebra (Carathèodory extension theorem). In particular, when the generators are open sets, these conditions are true and we can restrict the definition to the open sets alone. This example is going to be expanded further in Section 1.5.

Proposition 1.20 (Elementary properties of probability measure) *Let* $(\Omega, \mathscr{F}, \mathbf{P})$ *be a probability space. Then*

1. $\forall A, B \in \mathscr{F}$ *with* $A \subseteq B$, *then* $\mathbf{P}(A) \leq \mathbf{P}(B)$
2. $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B)$, $\forall A, B \in \mathscr{F}$.
3. *(General inclusion-exclusion formula, also named Poincaré formula):*

$$\mathbf{P}(A_1 \cup A_2 \cup \cdots \cup A_n) = \sum_{i=1}^{n} \mathbf{P}(A_i) - \sum_{i<j\leq n} \mathbf{P}(A_i \cap A_j) \tag{1.7}$$
$$+ \sum_{i<j<k\leq n} \mathbf{P}(A_i \cap A_j \cap A_k) - \cdots + (-1)^{n-1}\mathbf{P}(A_1 \cap A_2 \cap \cdots \cap A_n)$$

   Note that successive partial sums are alternating between over- and underestimating.

4. *(Finite subadditivity, sometimes called Boole's inequality):*

$$\mathbf{P}\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} \mathbf{P}(A_i), \quad \forall A_1, A_2, \ldots, A_n \in \mathscr{F}.$$

1.1.1 Null element of $\mathscr{F}$. Almost sure (a.s.) statements. Indicator of a set

An event $N \in \mathscr{F}$ is called a *null event* if $P(N) = 0$.

Definition 1.21 *A statement $\mathcal{S}$ about points $\omega \in \Omega$ is said to be true* almost surely *(a.s.), almost everywhere (a.e.), or with probability 1 (w.p.1) if the set M, defined as*

$$M := \{\omega \in \Omega | \, \mathcal{S}(\omega) \text{ is true}\},$$

is in $\mathscr{F}$ and $\mathbf{P}(M) = 1$, (or, equivalently M^c is a null set).

We will use the notions a.s., a.e., and w.p.1. to denote the same thing – the definition above. For example, we will say $X \geq 0$ a.s. and mean that $\mathbf{P}\{\omega | X(\omega) \geq 0\} = 1$, or equivalently $\mathbf{P}\{\omega | X(\omega) < 0\} = 0$. The notion of almost sure is a fundamental one in probability. Unlike in deterministic cases where something has to always be true no matter what, in probability we care about "the majority of the truth." In other words, probability recognizes that some phenomena may have extreme outcomes, but if they are extremely improbable then we do not care about them. Note that for this notion to make any sense, once again you have to think outside the realm of finite dimensional spaces. If the space Ω has a finite number of outcomes, say $10^{10000000000}$ all with a very small but strict positive probability, then neglecting any one of them makes the probability of any resulting set less than 1. To get what we are talking about with the null sets, you need once again to think about spaces with an infinite number of elements.

Definition 1.22 *We define the indicator function of an event A as the (simple) function $\mathbf{1}_A : \Omega \to \{0, 1\}$,*

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

Sometimes, this function is denoted as I_A.

Note that the indicator function is a regular function (not a set function). Indicator functions are very useful in probability theory. Here are some useful relationships:

$$\mathbf{1}_{\mathbf{A}\cap\mathbf{B}}(\omega) = 1_A(\omega)1_B(\omega), \quad \forall\, \omega \in \Omega$$

If the family of sets $\{A_i\}$ forms a partition of Ω (i.e., the sets A_i are disjoint and $\Omega = \bigcup_{i=1}^{n} A_i$), then

$$\mathbf{1}_B(\omega) = \sum_i 1_{B\cap A_i}(\omega), \text{ for any set } B \in \Omega$$

Remember this very simple function. We shall use it over and over throughout this book.

1.2 Conditional Probability

Let $(\Omega, \mathscr{F}, \mathbf{P})$ be a probability space.

Definition 1.23 *For $A, B \in \mathscr{F}$, with $\mathbf{P}(B) \neq 0$, we define the conditional probability of A given B by*

$$\mathbf{P}(A|\, B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)}.$$

We can immediately rewrite the formula above to obtain the *multiplicative rule*:

$$\begin{aligned}
&\mathbf{P}(A \cap B) = \mathbf{P}(A|\, B)\mathbf{P}(B),\\
&\mathbf{P}(A \cap B \cap C) = \mathbf{P}(A|\, B \cap C)\mathbf{P}(B|\, C)\mathbf{P}(C)\\
&\mathbf{P}(A \cap B \cap C \cap D) = \mathbf{P}(A|\, B \cap C \cap D)\mathbf{P}(B|\, C \cap D)\mathbf{P}(C|D)\mathbf{P}(D), \quad \text{and so on.}
\end{aligned}$$

This multiplicative rule is very useful for stochastic processes (part 2 of the book) and for estimation of parameters of a distribution.

Total probability formula: Given $A_1, A_2, \ldots, A_n$, a partition of Ω (i.e., the sets A_i are disjoint and $\Omega = \bigcup_{i=1}^{n} A_i$), then

$$\mathbf{P}(B) = \sum_{i=1}^{n} \mathbf{P}(B|\, A_i)\mathbf{P}(A_i), \quad \forall B \in \mathscr{F} \tag{1.8}$$

Bayes Formula: If $A_1, A_2, \ldots, A_n$ form a partition of Ω

$$\mathbf{P}(A_j|\, B) = \frac{\mathbf{P}(B|\, A_j)\mathbf{P}(A_j)}{\sum_{i=1}^{n} \mathbf{P}(B|\, A_i)\mathbf{P}(A_i)}, \quad \forall B \in \mathscr{F}\,. \tag{1.9}$$

EXAMPLE 1.5

A biker leaves the point O in the figure below. At each crossroad, the biker chooses a road at random. What is the probability that he arrives at point A ?

Let B_k, $k = 1, 2, 3, 4$ be the event that the biker passes through point B_k. These four events are mutually exclusive and they form a partition of the space. Moreover, they are equiprobable ($\mathbf{P}(B_k) = 1/4, \forall k \in \{1, 2, 3, 4\}$). Let A denote the event "the biker reaches the destination point A." Conditioned on each of the possible points B_1-B_4 of passing, we have

$$\begin{aligned}
\mathbf{P}(A|B_1) &= 1/4\\
\mathbf{P}(A|B_2) &= 1/2\\
\mathbf{P}(A|B_3) &= 1
\end{aligned}$$

At B_4 is slightly more complex. We have to use the multiplicative rule:

$$\begin{aligned}
\mathbf{P}(A|B_4) &= 1/4 + \mathbf{P}(A \cap B_5|B_4) + \mathbf{P}(A \cap B_6 \cap B_5|B_4)\\
&= 1/4 + \mathbf{P}(A|B_5 \cap B_4)\mathbf{P}(B_5|B_4)\\
&\quad + \mathbf{P}(A|B_6 \cap B_5 \cap B_4)\mathbf{P}(B_6|B_5 \cap B_4)\mathbf{P}(B_5|B_4)\\
&= 1/4 + 1/3(1/4) + 1(1/3)(1/4) = 3/12 + 2/12 = 5/12
\end{aligned}$$

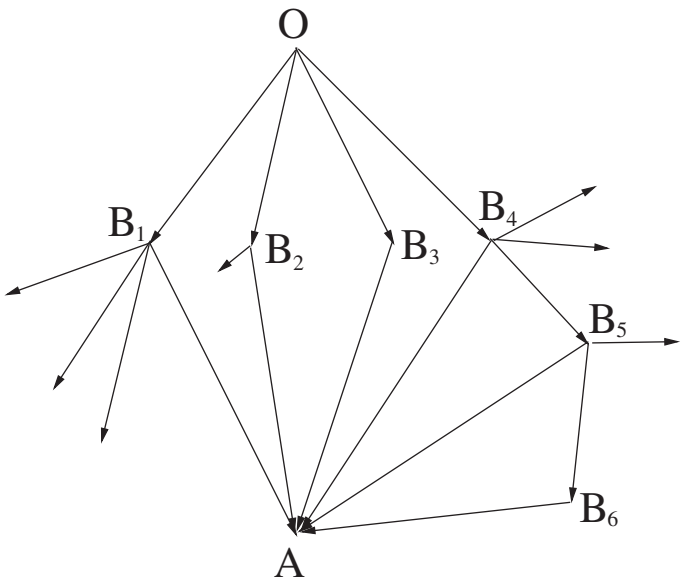

Figure 1.1 The possible trajectories of the biker. O is the origin point and A is the arrival point. B_k's are intermediate points.

Finally, by the law of total probability, we have

$$\begin{aligned}\mathbf{P}(A) &= \mathbf{P}(A|B_1)\mathbf{P}(B_1) + \mathbf{P}(A|B_2)\mathbf{P}(B_2) + \mathbf{P}(A|B_3)\mathbf{P}(B_3) + \mathbf{P}(A|B_4)\mathbf{P}(B_4) \\ &= 1/4(1/4) + 1/2(1/4) + 1/4(1) + 5/12(1/4) = 13/24\end{aligned}$$

■

EXAMPLE 1.6 De Mére's paradox

As a result of extensive observation of dice games, the French gambler Chevaliér De Mére noticed that the total number of spots showing on three dice thrown simultaneously turn out to be 11 more often than 12. However, from his point of view this is not possible since 11 occurs in six ways :

$(6:4:1); (6:3:2); (5:5:1); (5:4:2); (5:3:3); (4:4:3)$,

while 12 also occurs in six ways:

$(6:5:1); (6:4:2); (6:3:3); (5:5:2); (5:4:3); (4:4:4)$

What is the fallacy in the argument?

Proof (Solution due to Pascal): The argument would be correct if these "ways" would have the same probability. However, this is not true. For example, (6:4:1) occurs in 3! ways, (5:5:1) occurs in 3 ways, and (4:4:4) occurs in 1 way.

As a result, we can calculate $\mathbf{P}(11) = 27/216$; $\mathbf{P}(12) = 25/216$, and indeed his observation is correct and he should bet on 11 rather than on 12 if they have the same game payoff. ■

EXAMPLE 1.7 Another De Méré's paradox

Which one of the following is more probable?

1. Throw four dice and obtain at least one 6;
2. Throw two dice 24 time and obtain at least once a double 6.

Proof (Solution): For option 1: $1 - \mathbf{P}$(No 6)$= 1 - (5/6)^4 = 0.517747$.

For option 2: $1 - \mathbf{P}$(None of the 24 trials has a double 6)$= 1 - (35/36)^{24} = 0.491404$ ■

EXAMPLE 1.8 Monty Hall problem

This is a problem named after the host of the American television show "Let's make a deal." Simply put, at the end of a game you are left to choose between three closed doors. Two of them have nothing behind, and one contains a prize. You chose one door but the door is not opened automatically. Instead, the presenter opens another door that contains nothing. He then gives you the choice of changing the door or sticking with the initial choice.

Most people would say that it does not matter what you do at this time, but that is not true. In fact everything depends on the host's behavior. For example, if the host knows in advance where the prize is and always reveals at random some other door that does not contain anything, then it is always better to switch.

Proof (Solution): This problem generated a lot of controversy since its publication (in the 1970s) since the solution seems so counterintuitive. We mention articles talking about this problem in more detail (Morgan et al., 1991; Mueser and Granberg, 1991). We are presenting the problem here since it exemplifies the conditional probability reasoning. The key in any such problem is the sample space which has to be complete enough to be able to answer the questions asked.

Let D_i be the event that the prize is behind door i. Let SW be the event that switching wins the prize[2]

It does not matter which door we chose initially. The reasoning is identical with all three doors. So, we assume that initially we pick door 1.

Events D_i $i = 1, 2, 3$ are mutually exclusive, and we can write

$$\mathbf{P}(SW) = \mathbf{P}(SW|D_1)\mathbf{P}(D_1) + \mathbf{P}(SW|D_2)\mathbf{P}(D_2) + \mathbf{P}(SW|D_3)\mathbf{P}(D_3).$$

When the prize is behind door 1, since we chose door 1, the presenter has two choices for the door to show us. However, neither would contain the prize, and in

[2]As a side note, this event is the same as the event "not switching loses."

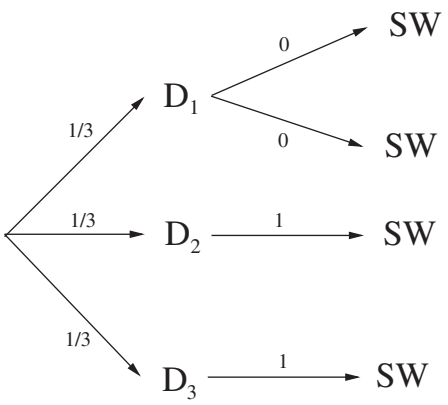

Figure 1.2 The tree diagram of conditional probabilities. Note that the presenter has two choices in case D_1, neither of which results in winning if switching the door.

either case switching does not result in winning the prize, therefore $\mathbf{P}(SW|D_1) = 0$. If the prize is behind door 2, since our choice is door 1, the presenter has no alternative but to show us the other door (3) which contains nothing. Thus, switching in this case results in winning the price. The same reasoning works if the prize is behind door 3. Therefore

$$\mathbf{P}(SW) = 1\frac{1}{3} + 1\frac{1}{3} + 0\frac{1}{3} = \frac{2}{3}$$

Thus switching has a higher probability of winning than not switching.

A generalization to n doors shows that it still is advantageous to switch, but the advantage decreases as $n \to \infty$. Specifically, in this case $\mathbf{P}(D_i) = 1/n$; $\mathbf{P}(SW|D_1) = 0$ still, but $\mathbf{P}(SW|D_i) = 1/(n-2)$ if $i \neq 1$, which gives

$$\mathbf{P}(SW) = \sum_{i=2}^{n} \frac{1}{n}\frac{1}{n-2} = \frac{n-1}{n-2}\frac{1}{n} > \frac{1}{n}$$

Furthermore, different presenter strategies produce different answers. For example, if the presenter offers the option to switch only when the player chooses the right door, then switching is always bad. If the presenter offers switching only when the player has chosen incorrectly, then switching always wins. These and other cases are analyzed in Rosenthal (2008). ■

EXAMPLE 1.9 Bertrand's box paradox

This problem was first formulated by Joseph Louis François Bertrand in his Calcul de Probabilités (Bertrand, 1889). In some sense this problem is related to the previous problem but it does not depend on any presenter strategy and the solution is much clearer. Solving this problem is an exercise in Bayes formula.

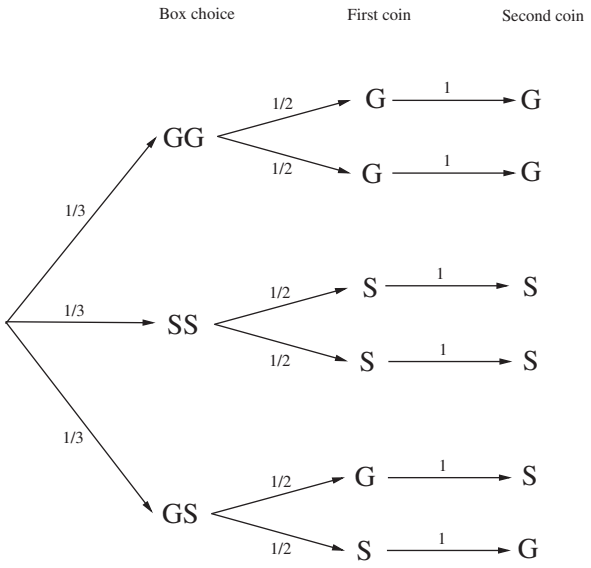

Figure 1.3 The tree diagram of conditional probabilities.

Suppose that we have three boxes. One box contains two gold coins, a second box has two silver coins, and a third box has one of each. We choose a box at random and from that box we choose a coin also at random. Then we look at the coin chosen. Given that the coin chosen was gold, what is the probability that the other coin in the box chosen is also gold. At first glance it may seem that this probability is $1/2$, but after calculation this probability turns out to be 2/3.

Proof (Solution):

We plot the sample space in Figure 1.3. Using this tree we can calculate the probability:

$$\mathbf{P}(\text{Second coin is } G|\text{First coin is } G) = \frac{\mathbf{P}(\text{Second coin is } G \text{ and First coin is } G)}{\mathbf{P}(\text{First coin is } G)}.$$

Now, using the probabilities from the tree we continue:

$$= \frac{\frac{1}{3}\frac{1}{2}1 + \frac{1}{3}\frac{1}{2}1}{\frac{1}{3}\frac{1}{2}1 + \frac{1}{3}\frac{1}{2}1 + \frac{1}{3}\frac{1}{2}1} = \frac{2}{3}.$$

Now that we have seen the solution, we can recognize a logical solution to the problem as well. Given that the coin seen is gold, we can throw away the middle box. If this would be box 1, then we have two possibilities that the other coin is gold (depending on which one we have chosen in the first place). If this is box 2, then there is one possibility (the remaining coin is silver). Thus the probability should be 2/3 since we have two out of three chances. Of course, this argument does not work if we do not choose the boxes with the same probability. ■

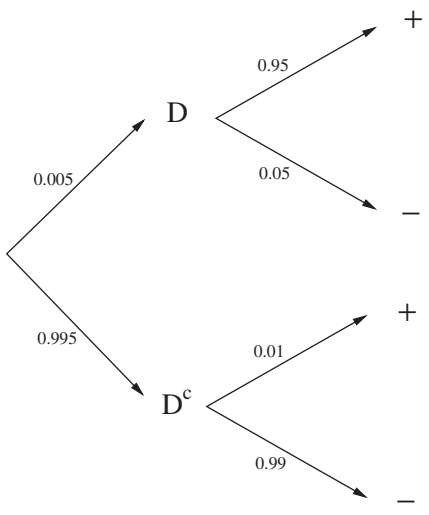

Figure 1.4 Blood test probability diagram

EXAMPLE 1.10

A blood test is 95% effective in detecting a certain disease when it is in fact present. However, the test yields a false positive result for 1% of the people who do not have the disease. If 0.5% of the population actually have the disease, what is the probability that a randomly chosen person is diseased, given that his test is positive?

Proof (Solution): This problem illustrates once again the application of the Bayes rule. I do not like to use the formula literally, instead I like to work from first principles and obtain the Bayes rule without memorizing anything. We start by describing the sample space. Refer to Figure 1.4 for this purpose.

So, given that the test is positive means that we have to calculate a conditional probability. We may write

$$\mathbf{P}(D|+) = \frac{\mathbf{P}(D \cap +)}{\mathbf{P}(+)} = \frac{\mathbf{P}(+|D)\mathbf{P}(D)}{\mathbf{P}(+)} = \frac{0.95(0.005)}{0.95(0.005) + 0.01(0.995)} = 0.323$$

How about if only 0.05% (i.e., 0.0005) of the population has the disease?

$$\mathbf{P}(D|+) = \frac{0.95(0.0005)}{0.95(0.0005) + 0.01(0.9995)} = 0.0454$$

This problem is an exercise in thinking. A good rate of correctly identifying the disease does not necessarily translate into a good rate of a person having the disease if the test is positive. The latter strongly depends on the actual proportion of the population having the disease. ■

EXAMPLE 1.11 Gambler's ruin problem

We conclude this section with an example which we will see many times throughout this book. This problem appeared in De Moivre's doctrine of chance, but an earlier version was also published by Huygens (1629–1695).

The formulation is simple: a game of heads or tails with a fair coin. Player wins 1 dollar if he successfully calls the side of the coin which lands upwards and loses \$1 otherwise. Suppose the initial capital is X dollars and he intends to play until he wins m dollars but no longer. What is the probability that the gambler will be ruined (loses all his/her money)?

Proof (Solution): We will display what is called a *first-step analysis*.

Let $p(x)$ denote the probability that the player is going to be eventually ruined if he starts with x dollars.

If he wins the next game, then he will have \$ $x+1$ and he will be ruined from this position with prob $p(x+1)$.

If he loses the next game, then he will have \$ $x-1$, so he is ruined from this position with prob $p(x-1)$.

Let R be the event in which he is eventually ruined. Let W be the event in which he wins the next trial. Let L be the event in which he loses this trial. Using the total probability formula, we get

$$\mathbf{P}(R) = \mathbf{P}(R|W)\mathbf{P}(W) + \mathbf{P}(R|L)\mathbf{P}(L) \Rightarrow p(x) = p(x+1)(1/2) + p(x-1)(1/2)$$

Is this true for all x? No. This is true for $x \geq 1$ and $x \leq m-1$. In the rest of the cases, we obviously have $p(0) = 1$ and $p(m) = 0$, which give the boundary conditions for the equation above.

This is a linear difference equation with constant coefficients. You can find a refresher of the general methodology in Appendix A.1.

Applying the method in our case gives the characteristic equation

$$y = \frac{1}{2}y^2 + \frac{1}{2} \Rightarrow y^2 - 2y + 1 = 0 \Rightarrow (y-1)^2 = 0 \Rightarrow y_1 = y_2 = 1$$

In our case the two solutions are equal, thus we seek a solution of the form $p(x) = (C + Dx)1^n = C + Dx$. Using the initial conditions, we get $p(0) = 1 \Rightarrow C = 1$ and $p(m) = 0 \Rightarrow C + Dm = 0 \Rightarrow D = -C/m = -1/m$; thus the general probability of ruin starting with wealth x is

$$p(x) = 1 - x/m.$$

■

1.3 Independence

Definition 1.24 *Two events, A and B, are independent if and only if*

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

The events $A_1, A_2, A_3, \ldots$ *are called* mutually independent *(or sometimes* simply independent*) if for every subset* J *of* $\{1, 2, 3, \ldots\}$ *we have*

$$\mathbf{P}\left(\bigcap_{j\in J} A_j\right) = \prod_{j\in J} \mathbf{P}(A_j)$$

The events $A_1, A_2, A_3, \ldots$ *are called* pairwise independent *(sometimes* jointly independent*) if*

$$\mathbf{P}\left(A_i \cap A_j\right) = \mathbf{P}(A_i)\mathbf{P}(A_j), \quad \forall i, j \in J.$$

Note that jointly independent does not imply independence.

Two sigma fields $\mathscr{G}, \mathscr{H} \in \mathscr{F}$ are $\mathbf{P}$-independent *if*

$$\mathbf{P}(G \cap H) = \mathbf{P}(G)\mathbf{P}(H), \quad \forall G \in \mathscr{G}, \forall H \in \mathscr{H}.$$

See Billingsley (1995) for the definition of independence of $k \geq 2$ *sigma algebras.*

EXAMPLE 1.12 The birthday problem

This is one of the oldest published probability problems. Suppose that there are n people meeting in a room. Each person is born on a certain day of the year and we assume that each day is equally likely for each individual (for simplicity we neglect leap years) and the probability of being born on a specific day is $1/365$. What is the probability that two or more people in the room share a birthday? What is the minimum number n such that this probability is greater than $1/2$?

This problem appeared for the first time in De Moivre's *The Doctrine of Chance: A method of calculating the probabilities of events in play* (1718).

The key is to calculate the complement of the probability requested. That is

$$\mathbf{P}\{\text{at least 2 people share a birthday}\} = 1 - \mathbf{P}\{\text{Nobody shares a birthday}\}$$

The probability that no one shares a birthday is easy to calculate using, for example, a conditional argument. To formalize using mathematical notation, let us denote the birthdays of the n individuals with $B_1, B_2, \ldots, B_n$. Let us denote the number of distinct elements in a set with $|\{B_1, B_2, \ldots, B_n\}|$. So, for example, $|\{B_1, B_2, \ldots, B_n\}| = n - 1$ means that there are $n - 1$ distinct elements in the set. Obviously, $|\{B_1, B_2, \ldots, B_n\}| = n$ means that there are no shared birthdays. Clearly, this probability is 0 if $n > 365$, so we assume that $n \leq 365$. We can write

$$\begin{aligned}\mathbf{P}\{|\{B_1, \ldots, B_n\}| = n\} &= \mathbf{P}\left\{|\{B_1, B_2, \ldots, B_n\}| = n \middle| |\{B_1, B_2, \ldots, B_{n-1}\}| = n-1\right\} \\ &\quad \mathbf{P}\{|\{B_1, B_2, \ldots, B_{n-1}\}| = n-1\} \\ &= \mathbf{P}\left\{B_n \notin \{B_1, \ldots, B_{n-1}\} \middle| |\{B_1, B_2, \ldots, B_{n-1}\}| = n-1\right\} \\ &\quad \mathbf{P}\{|\{B_1, B_2, \ldots, B_{n-1}\}| = n-1\}\end{aligned}$$

and we can follow the argument further until we get to B_1. The conditional probability is simply the probability that the nth individual is born on a different day

than the first $n-1$ given that the $n-1$ individuals all have different birthdays. This probability is $(365-(n-1))/365$ since there are $n-1$ choices which are not good. If we continue the argument until we reach the last individual, we finally obtain the desired probability:

$$\begin{aligned}\mathbf{P}\{\text{at least 2 people share the birthday}\} &= 1-\frac{365-(n-1)}{365}\cdots\frac{364}{365}\frac{365}{365}\\ &= 1-\frac{365!}{(365-n)!}\frac{1}{365^n}\end{aligned}$$

To answer the second question (what should n be to have the above probability over 0.5) we can use Stirling's approximation of the factorials appearing in the formula. This approximation will be given later, but to obtain the numbers at this time we are just going to run a search algorithm using a computer (n is at most 365).

Using a computer when $n=22$, we obtain the probability 0.475695308, and when $n=23$ we obtain the probability 0.507297234. So we need about 23 people in the room to have the probability of two or more people sharing the birthday greater than 0.5. However, once programmed properly it is easy to play around with a computer program. By changing the parameters, it is easy to see that with 30 people in the room the probability is over 0.7; with 35 people the probability is over 0.8; and if we have 41 people in a room we are 90% certain that at least two of them have a shared birthday.

1.4 Monotone Convergence Properties of Probability

Let us take a step back for a minute and comment on what we have seen thus far. σ-algebra differs from regular algebra in that it allows us to deal with a countable (not finite) number of sets. Again, this is a recurrent theme in probability: learning to deal with infinity. On finite spaces, things are more or less simple. One has to define the probability of each individual outcome and everything proceeds from there. However, even in these simple cases, imagine that one repeats an experiment (such as a coin toss) over and over. Again, we are forced to cope with infinity. This section introduces a way to deal with this infinity problem.

Let $(\Omega,\mathscr{F},\mathcal{P})$ be a probability space.

Lemma 1.25 *The following are true:*

1. *If $A_n, A\in\mathscr{F}$ and $A_n\uparrow A$ (i.e., $A_1\subseteq A_2\subseteq\ldots A_n\subseteq\ldots$ and $A=\bigcup_{n\geq 1}A_n$), then $\mathbf{P}(A_n)\uparrow\mathbf{P}(A)$ as a sequence of numbers.*

2. *If $A_n, A\in\mathscr{F}$ and $A_n\downarrow A$ (i.e., $A_1\supseteq A_2\supseteq\ldots A_n\supseteq\ldots$ and $A=\bigcap_{n\geq 1}A_n$), then $\mathbf{P}(A_n)\downarrow\mathbf{P}(A)$ as a sequence of numbers.*

3. *(Countable subadditivity) If* $A_1, A_2, \ldots,$ *and* $\bigcup_{i=1}^{\infty} A_n \in \mathcal{F}$, *with* A_i*'s not necessarily disjoint, then*

$$\mathbf{P}\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} \mathbf{P}(A_n)$$

Proof: 1. Let $B_1 = A_1, B_2 = A_2 \backslash A_1, \ldots, B_n = A_n \backslash A_{n-1}$. Because the sequence is increasing, we have that the B_i's are disjoint; thus

$$\mathbf{P}(A_n) = \mathbf{P}(B_1 \cup B_2 \cup \cdots \cup B_n) = \sum_{i=1}^{n} \mathbf{P}(B_i).$$

Thus, using countable additivity

$$\mathbf{P}\left(\bigcup_{n\geq 1} A_n\right) = \mathbf{P}\left(\bigcup_{n\geq 1} B_n\right) = \sum_{i=1}^{\infty} \mathbf{P}(B_i) = \lim_{n\to\infty} \sum_{i=1}^{n} \mathbf{P}(B_i) = \lim_{n\to\infty} \mathbf{P}(A_n).$$

2. Note that $A_n \downarrow A \;\Leftrightarrow\; {A_n}^c \uparrow A^c$ and from part 1 this means $1 - \mathbf{P}(A_n) \uparrow 1 - \mathbf{P}(A)$.

3. Let $B_1 = A_1, B_2 = A_1 \cup A_2, \ldots, B_n = A_1 \cup \cdots \cup A_n, \ldots$. From the finite subadditivity property in Proposition 1.20, we have that $\mathbf{P}(B_n) = \mathbf{P}(A_1 \cup \cdots \cup A_n) \leq \mathbf{P}(A_1) + \cdots + \mathbf{P}(A_n)$.

$\{B_n\}_{n\geq 1}$ is an increasing sequence of events, thus from part 1 we get that $\mathbf{P}(\bigcup_{n=1}^{\infty} B_n) = \lim_{n\to\infty} \mathbf{P}(B_n)$. Combining the two relations above, we obtain

$$\mathbf{P}(\bigcup_{n=1}^{\infty} A_n) = \mathbf{P}(\bigcup_{n=1}^{\infty} B_n) \leq \lim_{n\to\infty} (\mathbf{P}(A_1) + \cdots + \mathbf{P}(A_n)) = \sum_{n=1}^{\infty} \mathbf{P}(A_n)$$

■

Lemma 1.26 *The union of a countable number of* $\mathbf{P}$*-null sets is a* $\mathbf{P}$*-null set.*

This Lemma is a direct consequence of the countable subadditivity.

Recall from analysis: For a sequence of numbers $\{x_n\}_n$ $\limsup$ and $\liminf$ are defined

$$\limsup x_n = \inf_m \{\sup_{n\geq m} x_n\} = \lim_{m\to\infty} (\sup_{n\geq m} x_n),$$
$$\liminf x_n = \sup_m \{\inf_{n\geq m} x_n\} = \lim_{m\to\infty} (\inf_{n\geq m} x_n),$$

and they represent the highest (respectively lowest) limiting point of a subsequence included in $\{x_n\}_n$.

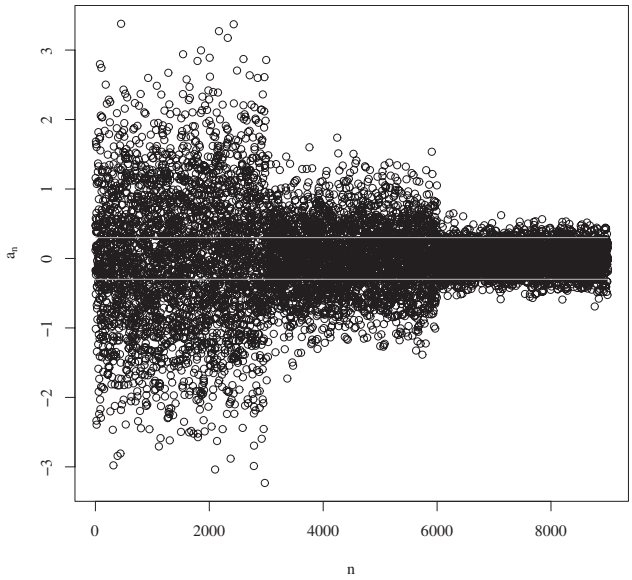

Figure 1.5 A sequence which has no limit. However, there exist sub-sequences that are convergent. The highest and lowest limiting points (gray lines) give the $\limsup$ and the $\liminf$, respectively

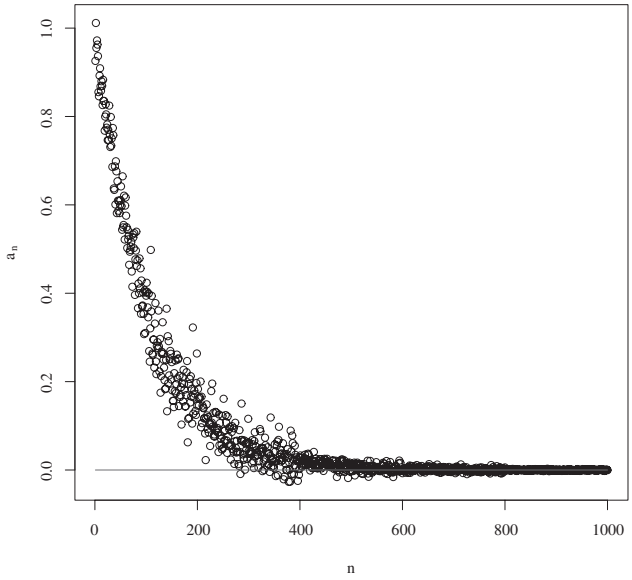

Figure 1.6 A sequence which has a limit (gray line). The limiting point is equal to both the $\limsup$ and the $\liminf$ of the sequence

Note that if z is a number such that $z > \limsup x_n$, then $x_n < z$ eventually[3].
Likewise, if $z < \limsup x_n$, then $x_n > z$ infinitely often[4].
These notions are translated to probability in the following way.

Definition 1.27 *Let $A_1, A_2, \ldots$ be an infinite sequence of events in some probability space $(\Omega, \mathscr{F}, \mathbf{P})$. We define the following events:*

$$\begin{aligned}\limsup_{n\to\infty} A_n &= \bigcap_{n\geq 1} \bigcup_{m=n}^{\infty} A_m = \{\omega : \omega \in A_n \textit{ for infinitely many } n\} \\ &= \{A_n \textit{ infinitely often}\} \\ \liminf_{n\to\infty} A_n &= \bigcup_{n\geq 1} \bigcap_{m=n}^{\infty} A_m = \{\omega : \omega \in A_n \textit{ for all } n \textit{ large enough}\} \\ &= \{A_n \textit{ eventually}\}\end{aligned}$$

Let us clarify the notions of "infinitely often" and "eventually" a bit more. We say that an outcome ω happens infinitely often for the sequence $A_1, A_2, \ldots, A_n, \ldots$ if ω is in the set $\bigcap_{n=1}^{\infty} \bigcup_{m\geq n} A_m$. This means that for any n (no matter how big) there exist an $m \geq n$ and $\omega \in A_m$.

We say that an outcome ω happens eventually for the sequence $A_1, A_2, \ldots, A_n, \ldots$ if ω is in the set $\bigcup_{n=1}^{\infty} \bigcap_{m\geq n} A_m$. This means that there exists an n such that for all $m \geq n$, $\omega \in A_m$, so from this particular n and up, ω is in all the sets.

Why do we give such complicated definitions? The basic intuition is the following: say, you roll a die infinitely many times, then it is obvious what it means for the outcome 1 to appear infinitely often. Also, we can say that the average of the rolls will eventually be arbitrarily close to 3.5 (this will be shown later). It is not very clear how to put this in terms of events happening infinitely often. The framework above provides a generalization to these notions.

The Borel Cantelli lemmas

With these definitions, we are now capable of giving two important lemmas.

Lemma 1.28 (First Borel–Cantelli) *If $A_1, A_2, \ldots$ is any infinite sequence of events with the property $\sum_{n\geq 1} \mathbf{P}(A_n) < \infty$, then*

$$\mathbf{P}\left(\bigcap_{n=1}^{\infty} \bigcup_{m\geq n} A_m\right) = \mathbf{P}\left(A_n \textit{ events are true infinitely often}\right) = 0$$

This lemma essentially says that if the probabilities of events go to zero and the sum is convergent, then necessarily A_n will stop occurring. However, the reverse of

[3] i.e., there is some n_0 very large so that $x_n < z$, for all $n \geq n_0$
[4] That is, for any n there exists an $m \geq n$ such that $x_m > z$.

the statement is not true. To make it hold, we need a very strong condition (independence).

Lemma 1.29 (Second Borel–Cantelli) *If $A_1, A_2, \ldots$ is an infinite sequence of* ***independent*** *events, then*

$$\sum_{n\geq 1} \mathbf{P}(A_n) = \infty \quad \Leftrightarrow \quad \mathbf{P}(A_n \textit{ infinitely often}) = 1.$$

Proof:

First Borel–Cantelli

$$\mathbf{P}\left(A_n \text{ i.o.}\right) = \mathbf{P}\left(\bigcap_{n\geq 1}\bigcup_{m=n}^{\infty} A_m\right) \leq \mathbf{P}\left(\bigcup_{n=m}^{\infty} A_m\right) \leq \sum_{m=n}^{\infty} \mathbf{P}(A_m), \forall n,$$

where we used the definition and countable subadditivity. By the hypothesis, the sum on the right is the tail end of a convergent series and therefore converges to zero as $n \to \infty$. Thus we are done. ∎

Proof (Second Borel–Cantelli:): The "⇒" part. Clearly, showing that $\mathbf{P}\left(A_n \text{ i.o.}\right) = \mathbf{P}(\limsup A_n) = 1$ is the same as showing that $\mathbf{P}\left((\limsup A_n)^c\right) = 0$.

By the definition of $\limsup$ and the DeMorgan's laws,

$$(\limsup A_n)^c = \left(\bigcap_{n\geq 1}\bigcup_{m=n}^{\infty} A_m\right)^c = \bigcup_{n\geq 1}\bigcap_{m=n}^{\infty} A_m^c.$$

Therefore, it is enough to show that $\mathbf{P}(\bigcap_{m=n}^{\infty} A_m^c) = 0$ for all n (recall that a countable union of null sets is a null set). However

$$\begin{aligned}
\mathbf{P}\left(\bigcap_{m=n}^{\infty} A_m^c\right) &= \lim_{r\to\infty} \mathbf{P}\left(\bigcap_{m=n}^{r} A_m^c\right) = \underbrace{\lim_{r\to\infty} \prod_{m=n}^{\infty} \mathbf{P}\left(A_m^c\right)}_{\text{by independence}} \\
&= \lim_{r\to\infty} \prod_{m=n}^{r} \left(1 - \mathbf{P}(A_m)\right) \leq \underbrace{\lim_{r\to\infty} \prod_{m=n}^{r} e^{-\mathbf{P}(A_m)}}_{1-x\leq e^{-x} \text{ if } x\geq 0} \\
&= \lim_{r\to\infty} e^{-\sum_{m=n}^{r} \mathbf{P}(A_m)} = e^{-\sum_{m=n}^{\infty} \mathbf{P}(A_m)} = 0
\end{aligned}$$

The last equality follows since $\sum \mathbf{P}(A_n) = \infty$.
Note that we have used the inequality $1 - x \leq e^{-x}$, which is true if $x \in [0, \infty)$. One can prove this inequality with elementary analysis.

The "⇐" part. This implication is the same as the first lemma. Indeed, assume by absurd that $\sum \mathbf{P}(A_n) < \infty$. By the first Borel–Cantelli lemma this implies that $\mathbf{P}(A_n \text{ i.o.}) = 0$, a contradiction with the hypothesis of the implication. ∎

The Fatou lemmas

Again, assume that $A_1, A_2, \ldots$ is a sequence of events.

Lemma 1.30 (Fatou lemma for sets) *Given any measure (not necessarily finite) μ, we have*

$$\mu(A_n \textit{ eventually}) = \mu(\liminf_{n\to\infty} A_n) \leq \liminf_{n\to\infty} \mu(A_n)$$

Proof: Recall that $\liminf_{n\to\infty} A_n = \bigcup_{n\geq 1} \bigcap_{m=n}^{\infty} A_m$, and denote this set with A. Let $B_n = \bigcap_{m=n}^{\infty} A_m$, which is an increasing sequence (less intersections as n increases) and $B_n \uparrow A =$. By the monotone convergence property of measure (Lemma 1.25), $\mu(B_n) \to \mu(A)$. However,

$$\mu(B_n) = \mu(\bigcap_{m=n}^{\infty} A_m) \leq \mu(A_m), \forall m \geq n,$$

thus $\mu(B_n) \leq \inf_{m\geq n} \mu(A_m)$. Therefore

$$\mu(A) \leq lim_{n\to\infty} \inf_{m\geq n} \mu(A_m) = \liminf_{n\to\infty} \mu(A_n)$$

■

Lemma 1.31 (The reverse of the Fatou lemma) *If $\mathbf{P}$ is a finite measure (e.g., probability measure), then*

$$\mathbf{P}(A_n \textit{ i.o.}) = \mathbf{P}(\limsup_{n\to\infty} A_n) \geq \limsup_{n\to\infty} \mathbf{P}(A_n).$$

Proof: This proof is entirely similar. Recall that $\limsup_{n\to\infty} A_n = \bigcap_{n\geq 1} \bigcup_{m=n}^{\infty} A_m$, and denote this set with A. Let $B_n = \bigcup_{m=n}^{\infty} A_m$. Then clearly B_n is a decreasing sequence and $B_n \downarrow A$. By the monotone convergence property of measure (Lemma 1.25) and since the measure is finite, $\mathbf{P}(B_1) < \infty$ so $\mathbf{P}(B_n) \to \mathbf{P}(A)$. However,

$$\mathbf{P}(B_n) = \mathbf{P}(\bigcup_{m=n}^{\infty} A_m) \geq \mathbf{P}(A_m), \forall m \geq n,$$

thus $\mathbf{P}(B_n) \geq \sup_{m\geq n} \mathbf{P}(A_m)$, again since the measure is finite. Therefore

$$\mathbf{P}(A) \geq lim_{n\to\infty} \sup_{m\geq n} \mathbf{P}(A_m) = \limsup_{n\to\infty} \mathbf{P}(A_n)$$

■

Kolmogorov zero-one law

We like to present this theorem since it introduces the concept of *a sequence of σ-algebras*, a notion essential for stochastic processes.

For a sequence $A_1, A_2, \dots$ of events in the probability space $(\Omega, \mathscr{F}, \mathscr{P})$, consider the generated sigma algebras $\mathscr{T}_n = \sigma(A_n, A_{n+1}, \dots)$ and their intersection

$$\mathscr{T} = \bigcap_{n=1}^{\infty} \mathscr{T}_n = \bigcap_{n=1}^{\infty} \sigma(A_n, A_{n+1}, \dots),$$

called the tail σ-field.

Theorem 1.32 (Kolmogorov's 0–1 law) *If the sets $A_1, A_2, \dots$ are independent in the sense of definition 1.24, then every event A in the tail σ field $\mathscr{T}$ defined above has probability $\mathbf{P}(A)$ either 0 or 1.*

Remark 1.33 *This theorem says that any event in the tail sigma algebra either happens all the time or it does not happen at all. As any Kolmogorov result, this one is very useful in practice. In many applications, we want to show that, despite the random nature of the phenomenon, eventually something will happen for sure. One needs to show that the event desired is in the tail sigma algebra and give an example (a sample path) where the limiting behavior is observed. However, this is an old result. A practical limitation today is the (very strong) assumption that the sets are independent.*

Proof: We skip this proof and only give the steps of the theorem. The idea is to show that A is independent of itself, thus $\mathbf{P}(A \cap A) = \mathbf{P}(A)\mathbf{P}(A) \Rightarrow \mathbf{P}(A) = \mathbf{P}(A)^2 \Rightarrow \mathbf{P}(A)$ is either 0 or 1. The steps of this proof are as follows:

1. First define $\mathscr{A}_n = \sigma(A_1, \dots, A_n)$ and show that it is independent of $\mathscr{T}_{n+1}$ for all n.
2. Since $\mathscr{T} \subseteq \mathscr{T}_{n+1}$ and $\mathscr{A}_n$ is independent of $\mathscr{T}_{n+1}$, then $\mathscr{A}_n$ and $\mathscr{T}$ are independent for all n.
3. Define $\mathscr{A}_\infty = \sigma(A_1, A_2, \dots)$. Then from the previous step we deduce that $\mathscr{A}_\infty$ and $\mathscr{T}$ are independent.
4. Finally, since $\mathscr{T} \subseteq \mathscr{A}_\infty$ by the previous step, $\mathscr{T}$ is independent of itself and the result follows.

■

Note that $\limsup A_n$ and $\liminf A_n$ are tail events. However, it is only in the case when the original events are independent that we can apply Kolmogorov's theorem. Thus in that case $\mathbf{P}\{A_n \text{ i.o.}\}$ is either 0 or 1.

1.5 Lebesgue Measure on the Unit Interval (0,1]

We conclude this chapter with the most important measure available. This is the unique measure that makes things behave in a normal way (e.g., the interval $(0.2, 0.5)$ has measure 0.3).

Let $\Omega = (0, 1]$. Let $\mathscr{F}_0$ = class of semiopen subintervals (a,b] of Ω. For an interval $I = (a, b] \in \mathscr{F}_0$, define $\lambda(I) = |I| = b - a$. Let $\varnothing \in \mathscr{F}_0$ the element of length 0. Let $\mathscr{B}_0$=the algebra of finite disjoint unions of intervals in (0,1]. This algebra is not a σ-algebra. The proof is in Problem 1.4 at the end of this chapter.

If $A = \bigcup_{i=1}^{n} I_n \in \mathscr{B}_0$ with I_n disjoint $\mathscr{F}_0$ sets, then

$$\lambda(A) = \sum_{i=1}^{n} \lambda(I_i) = \sum_{i=1}^{n} |I_i|$$

The goal is to show that λ is countably additive on the algebra $\mathscr{B}_0$. This will allow us to construct a measure (actually a probability measure since we are working on (0,1]) using the Caratheodory's theorem (Theorem 1.36). The resulting measure is well defined and is called the *Lebesgue measure*.

Theorem 1.34 (Theorem for the length of intervals) *Let $I = (a, b] \subseteq (0, 1]$ and I_k of the form $(a_k, b_k]$ bounded but not necessarily in $(0, 1]$.*

(i) If $\bigcup_k I_k \subseteq I$ and I_k are disjoint, then $\sum_k |I_k| \leq |I|$.

(ii) If $I \subseteq \bigcup_k I_k$ (with the I_k not necessarily disjoint), then $|I| \leq \sum_k |I_k|$.

(iii) If $I = \bigcup_k I_k$ and I_k disjoint, then $|I| = \sum_k |I_k|$.

Proof: Exercise (*Hint:* use induction) ∎

Note: Part (iii) shows that the function λ is well defined.

Theorem 1.35 *λ is a (countably additive) probability measure on the field $\mathscr{B}_0$. λ is called the Lebesgue measure restricted to the algebra $\mathscr{B}_0$*

Proof: Let $A = \bigcup_{k=1}^{\infty} A_k$, where A_k are disjoint $\mathscr{B}_0$ sets. By definition of $\mathscr{B}_0$,

$$A_k = \bigcup_{j=1}^{m_k} J_{k_j}, \quad A = \bigcup_{i=1}^{n} I_i,$$

where the J_{k_j} are disjoint. Then,

$$\lambda(A) = \sum_{i=1}^{n} |I_i| = \sum_{i=1}^{n} (\sum_{k=1}^{\infty} \sum_{j=1}^{m_k} |I_i \cap J_{k_j}|) = \sum_{k=1}^{\infty} \sum_{j=1}^{m_k} (\sum_{i=1}^{n} |I_i \cap J_{k_j}|)$$

and since $A \cap J_{k_j} = J_{k_j} \Rightarrow |A \cap J_{k_j}| = \sum_{i=1}^{n} |I_i \cap J_{k_j}| = |J_{k_j}|$, the above is continued:

$$= \sum_{k=1}^{\infty} \underbrace{\sum_{j=1}^{m_k} |J_{k_j}|}_{=|A_k|} = \sum_{k=1}^{\infty} \lambda(A_k)$$

∎

The next theorem will extend the Lebesgue measure to the whole $(0,1]$, thus we define the probability space $((0,1],\mathscr{B}((0,1]),\lambda)$. The same construction with minor modifications works in $(\mathbb{R},\mathscr{B}(\mathbb{R}),\lambda)$.

Theorem 1.36 (Caratheodory's extension theorem) *A probability measure on an algebra has a unique extension to the generated σ-algebra.*

Note: The Caratheodory theorem practically constructs all the interesting probability models. However, once we construct our models we have no further need of the theorem. It also reminds us of the central idea in the theory of probabilities: If one wants to prove something for a big set, one needs to look first at the generators of that set.

Proof: (skipped), in the problems 1.11,1.12. ∎

Definition 1.37 (Monotone class) *A class $\mathscr{M}$ of subsets in Ω* is monotone *if it is closed under the formation of monotone unions and intersections, that is:*

(i) $A_1, A_2, \cdots \in \mathscr{M}$ *and* $A_n \subset A_{n+1}, \bigcup_n A_n = A \Rightarrow A \in \mathscr{M}$,

(ii) $A_1, A_2, \cdots \in \mathscr{M}$ *and* $A_n \supset A_{n+1} \Rightarrow \bigcap_n A_n \in \mathscr{M}$.

The next theorem is only needed for the proof of the Caratheodory theorem. However, the proof is interesting and thus is presented here.

Theorem 1.38 *If $\mathscr{F}_0$ is an algebra and $\mathscr{M}$ is a monotone class, then $\mathscr{F}_0 \subseteq \mathscr{M} \Rightarrow \sigma(\mathscr{F}_0) \subseteq \mathscr{M}$.*

Proof: Let $m(\mathscr{F}_0)$ = minimal monotone class over $\mathscr{F}_0$ = the intersection of all monotone classes containing $\mathscr{F}_0$

We will prove that $\sigma(\mathscr{F}_0) \subseteq m(\mathscr{F}_0)$.

To show this, it is enough to prove that $m(\mathscr{F}_0)$ is an algebra. Then Exercise 1.12 will show that $m(\mathscr{F}_0)$ is a σ algebra. Since $\sigma(\mathscr{F}_0)$ is the smallest, the conclusion follows.

To this end, let $\mathscr{G} = \{A : A^c \in m(\mathscr{F}_0)\}$.

(i) Since $m(\mathscr{F}_0)$ is a monotone class, so is $\mathscr{G}$.

(ii) Since $\mathscr{F}_0$ is an algebra, its elements are in $\mathscr{G} \Rightarrow \mathscr{F}_0 \subset \mathscr{G}$.

(i) and (ii) $\Rightarrow m(\mathscr{F}_0) \subseteq \mathscr{G}$. Thus $m(\mathscr{F}_0)$ is closed under complementarity.

Now define $\mathscr{G}_1 = \{A : A \cup B \in m(\mathscr{F}_0), \forall B \in \mathscr{F}_0\}$.

We show that $\mathscr{G}_1$ is a monotone class:

Let $A_n \nearrow$ an increasing sequence of sets, $A_n \in \mathscr{G}_1$. By definition of $\mathscr{G}_1$, for all n $A_n \cup B \in m(\mathscr{F}_0), \forall B \in \mathscr{F}_0$.

But $A_n \cup B \supseteq A_{n-1} \cup B$, and thus the definition of $m(\mathscr{F}_0)$ implies

$$\bigcup_n (A_n \cup B) \in m(\mathscr{F}_0), \forall B \in \mathscr{F}_0 \Rightarrow \left(\bigcup_n A_n\right) \cup B \in m(\mathscr{F}_0), \forall B,$$

and thus $\bigcup_n A_n \in \mathscr{G}_1$.

This shows that $\mathscr{G}_1$ is a monotone class. But since $\mathscr{F}_0$ is an algebra, its elements (the contained sets) are in $\mathscr{G}_1$[5], thus $\mathscr{F}_0 \subset \mathscr{G}_1$. Since $m(\mathscr{F}_0)$ is the smallest monotone class containing $\mathscr{F}_0$, we immediately have $m(\mathscr{F}_0) \subseteq \mathscr{G}_1$.

Let $\mathscr{G}_2 = \{B : A \cup B \in m(\mathscr{F}_0), \forall A \in m(\mathscr{F}_0)\}$

$\mathscr{G}_2$ **is a monotone class.** (identical proof – see problem 1.11)

Let $B \in \mathscr{F}_0$. Since $m(\mathscr{F}_0) \subseteq \mathscr{G}_1$ for any set $A \in m(\mathscr{F}_0) \Rightarrow A \cup B \in m(\mathscr{F}_0)$. Thus, by the definition of $\mathscr{G}_2 \Rightarrow B \in \mathscr{G}_2 \Rightarrow \mathscr{F}_0 \subseteq \mathscr{G}_2$.

The previous implication and the fact that $\mathscr{G}_2$ is a monotone class imply that $m(\mathscr{F}_0) \subseteq \mathscr{G}_2$.

Therefore, $\forall A, B \in m(\mathscr{F}_0) \Rightarrow A \cup B \in m(\mathscr{F}_0) \Rightarrow m(\mathscr{F}_0)$ is an algebra. ∎

Problems

1.1 Roll a die. Then $\Omega = \{1, 2, 3, 4, 5, 6\}$. An example of an event is $A = \{$ Roll an even number$\} = \{2, 4, 6\}$. Find the cardinality (number of elements) of $\mathscr{P}(\Omega)$ in this case.

1.2 Suppose $\Omega = \{a, b, c\}$ is a probability space, and a discrete probability function is introduced such that $\mathbf{P}(\{a\}) = 1/3$ and $\mathbf{P}(\{b\}) = 1/2$. List all events in the maximal sigma algebra $\mathcal{P}(\Omega)$ and calculate the probability of each such event.

1.3 Suppose two events A and B are in some space Ω. List the elements of the generated σ algebra $\sigma(A, B)$ in the following cases:

a) $A \cap B = \emptyset$ b) $A \subset B$

c) $A \cap B \neq \emptyset$; $A \setminus B \neq \emptyset$ and $B \setminus A \neq \emptyset$.

1.4 An algebra which is not a σ-algebra

Let $\mathscr{B}_0$ be the collection of sets of the form $(a_1, a_1'] \cup (a_2, a_2'] \cup \cdots \cup (a_m, a_m']$, for any $m \in \mathbb{N}^* = \{1, 2 \ldots\}$ and all $a_1 < a_1' < a_2 < a_2' < \cdots < a_m < a_m'$ in $\Omega = (0, 1]$.

Verify that $\mathscr{B}_0$ is an algebra. Show that $\mathscr{B}_0$ is not a σ-algebra.

1.5 Let $\mathscr{F} = \{A \subseteq \Omega | \, A$ finite **or** A^c is finite$\}$.

a) Show that $\mathscr{F}$ is an algebra.

b) Show that, if Ω is finite, then $\mathscr{F}$ is a σ-algebra.

c) Show that, if Ω is infinite, then $\mathscr{F}$ is **not** a σ-algebra.

[5]One can just verify the definition of $\mathscr{G}_1$ for this.

1.6 A σ-Algebra does not necessarily contain all the events in Ω
Let $\mathscr{F} = \{A \subseteq \Omega | \, A \text{ countable } \textbf{or } A^c \text{ is countable}\}$. Show that $\mathscr{F}$ is a σ-algebra. Note that, if Ω is uncountable, it implies that it contains a set A such that both A and A^c are uncountable and thus $A \notin \mathscr{F}$.

1.7 Show that the Borel sets of $\mathbb{R}$ $\mathscr{B} = \sigma(\{(-\infty, x] | \, x \in \mathbb{R}\})$.

Hint: show that the generating set is the same, that is, show that any set of the form $(-\infty, x]$ can be written as countable union (or intersection) of open intervals, and, vice versa, that any open interval in $\mathbb{R}$ can be written as a countable union (or intersection) of sets of the form $(-\infty, x]$.

1.8 Show that the following classes all generate the Borel σ-algebra, or, put differently, show the equality of the following collections of sets:

$$\begin{aligned}\sigma((a,b) : a < b \in \mathbb{R}) = \sigma([a,b] : a < b \in \mathbb{R}) = \sigma((-\infty, b) : b \in \mathbb{R}) \\ = \sigma((-\infty, b) : b \in \mathbb{Q}),\end{aligned}$$

where $\mathbb{Q}$ is the set of rational numbers.

1.9 Properties of probability measures
Prove properties 1–4 in the Proposition 1.20 on page 21.

Hint: You only have to use the definition of probability. The only thing nontrivial in the definition is the countable additivity property.

1.10 No matter how many zeros do not add to more than zero
Prove the Lemma 1.26 on page 32.
Hint: You may use countable subadditivity.

1.11 If $\mathscr{F}_0$ is an algebra, $m(\mathscr{F}_0)$ is the minimal monotone class over $\mathscr{F}_0$, and $\mathscr{G}_2$ is defined as

$$\mathscr{G}_2 = \{B : A \cup B \in m(\mathscr{F}_0), \forall A \in m(\mathscr{F}_0)\}$$

Then show that $\mathscr{G}_2$ is a monotone class.
Hint: Look at the proof of theorem 1.38 on page 39, and repeat the arguments therein.

1.12 A monotone algebra is a σ-algebra
Let $\mathscr{F}$ be an algebra that is also a monotone class. Show that $\mathscr{F}$ is a σ-algebra.

1.13 Prove the *total probability formula* equation (1.8) and the *Bayes Formula* equation 1.9.

1.14 If two events are such that $A \cap B = \emptyset$, are A and B independent? Justify.

1.15 Show that $\mathbf{P}(A|B) = \mathbf{P}(A)$ is the same as independence of the events A and B.

1.16 Prove that, if two events A and B are independent, then so are their complements.

1.17 Generalize the previous problem to n sets using induction.

1.18 Calculating Odds. If we know the probabilities of a certain event, another way we can express these probabilities is through the odds – that is, the return given to betting on the event. To give an example, we can easily calculate the probability of the next card in a deck of 52 cards being a red card. So the odds offered on the next card being a red card are 1:1; that is, for every dollar bet, I give back 1 more dollar if the bet is a win. The probability is $0.5 = \frac{1}{1+1}$. In the same spirit, the odds given to draw a club are 3:1 or 3$ for every 1$ bet. (In fact, the return is 4 if you count the dollar initially bet.) The probability of course is $0.25 = \frac{1}{1+3}$. This idea will be seen in much more detail later when we talk about martingales. What happens if the probabilities are not of the form $1/n$. For example, it is easy to calculate that the probability of drawing a black ball from an urn containing three red and four black balls is $4/7$. No problem, the correct odds for betting on black are 3:4; that is, for every four dollars bet, you receive back three more. That is easy to see because the event has probability greater than $1/2$. So the next questions are all asking you to formulate the correct odds (in a fair game) of the following events:

a) A card chosen at random from a deck of 52 is an ace.
b) A card chosen at random is either a Queen or a Club.
c) In 4 cards drawn from a deck of 52, there are at least 2 Hearts.
d) A hand of bridge (13 cards) is missing a suit (i.e., does not contain all four clubs, hearts, diamond, and spades).
e) You see two heads when tossing a fair coin twice.
f) You see either a sum of 7 or a sum of 11 when rolling two six-sided dies.

Of course, you can also reconstruct the probabilities from the odds. What are the probabilities of the following events?

g) You are offered 3:1 odds that the Lakers will win the next game with the Pacers.
h) Odds 5:2 that the last shot is taken by player X from Lakers.
i) Odds 30:1 that the game is won on a last shot from player Y from the Pacers.

You should see by now from this very simple example that, in order that the probabilities to be consistent, one must have some sort of a model to avoid arbitrage (i.e., making money with no risk).

1.19 One urn contains w_1 white balls and b_1 black balls. Another urn contains w_2 white balls and b_2 black balls. A ball is drawn at random from each urn, then one of the two such chosen is selected at random.

a) What is the probability that the final ball selected is white?
b) Given that the final ball selected was white, what is the probability that in fact it came from the first urn (with w_1 and b_1 balls).

1.20 At the end of a well-known course, the final grade is decided with the help of an oral examination. There are a total of m possible subjects listed on some pieces of paper. Of them, n are generally considered "easy."

Each student enrolled in the class, one after another, draws a subject at random, and then presents it. Of the first two students, who has the better chance of drawing a "favorable" subject?

1.21 Andre Agassi and Pete Sampras decide to play a number of games together. They play nonstop and at the end it turns out that Sampras won n games while Agassi m, where $n > m$. Assume that in fact any possible sequence of games was possible to reach this result. Let $P_{n,m}$ denote the probability that from the first game until the last Sampras is always in the lead. Find

1. $P_{2,1}$; $P_{3,1}$; $P_{n,1}$
2. $P_{3,2}$; $P_{4,2}$; $P_{n,2}$
3. $P_{4,3}$; $P_{5,3}$; $P_{5,4}$
4. Make a conjecture about a formula for $P_{n,m}$.

1.22 My friend Andrei has designed a system to win at the roulette. He likes to bet on red, but he waits until there have been six previous black spins and only then he bets on red. He reasons that the chance of winning is quite large since the probability of seven consecutive back spins is quite small. What do you think of his system. Calculate the probability that he wins using this strategy.

Actually, Andrei plays his strategy four times and he actually wins three times out of the four he played. Calculate the probability of the event that just occurred.

1.23 Ali Baba is caught by the sultan while stealing his daughter. The sultan is being gentle with him and he offers Ali Baba a chance to regain his liberty.
There are two urns and m white balls and n black balls. Ali Baba has to put the balls in the two urns; however, he likes the only condition that no urn is empty. After that, the sultan will chose an urn at random, and then pick a ball from that urn. If the chosen ball is white Ali Baba is free to go, otherwise Ali Baba's head will be at the same level as his legs.
How should Ali Baba divide the balls to maximize his chance of survival?