CAMBRIDGE
Cambridge University Press & Assessment
978-1-108-48967-6 — Principles of Statistical Analysis
Ery Arias-Castro
Excerpt
More Information

Part I

Elements of Probability Theory

1

Axioms of Probability Theory

Probability theory is the branch of mathematics that models and studies random phenomena. Although randomness has been the object of much interest over many centuries, the theory only reached maturity with *Kolmogorov's axioms*[1] in the 1930s [195].

As a mathematical theory founded on Kolmogorov's axioms, *Probability Theory* is essentially uncontroversial at this point. However, the notion of probability (i.e., chance) remains somewhat controversial. We will adopt here the frequentist notion of probability [193], which defines the chance that a particular experiment results in a given outcome as the limiting frequency of this event as the experiment is repeated an increasing number of times. The problem of giving probability a proper definition as it concerns real phenomena is discussed in [67] (with a good dose of humor).

1.1 Elements of Set Theory

Kolmogorov's formalization of probability relies on some basic notions of *Set Theory*.

A *set* is simply an abstract collection of 'objects', sometimes called *elements* or *items*. Let Ω denote such a set. A *subset* of Ω is a set made of elements that belong to Ω. In what follows, a set will be a subset of Ω.

We write $\omega \in \mathcal{A}$ when the element ω belongs to the set $\mathcal{A}$. And we write $\mathcal{A} \subset \mathcal{B}$ when set $\mathcal{A}$ is a subset of set $\mathcal{B}$. This means that $\omega \in \mathcal{A} \Rightarrow \omega \in \mathcal{B}$. A set with only one element ω is denoted $\{\omega\}$ and is called a *singleton*. Note that $\omega \in \mathcal{A} \Leftrightarrow \{\omega\} \subset \mathcal{A}$. The *empty set* is defined as a set with no elements and is denoted $\emptyset$. By convention, it is included in any other set.

Problem 1.1 Prove that $\subset$ is transitive, meaning that if $\mathcal{A} \subset \mathcal{B}$ and $\mathcal{B} \subset \mathcal{C}$, then $\mathcal{A} \subset \mathcal{C}$.

[1] Named after Andrey Kolmogorov (1903–1987).

The following are some basic set operations.

- *Intersection and disjointness* The intersection of two sets $\mathcal{A}$ and $\mathcal{B}$ is the set with all the elements belonging to both $\mathcal{A}$ and $\mathcal{B}$, and is denoted $\mathcal{A} \cap \mathcal{B}$. $\mathcal{A}$ and $\mathcal{B}$ are said to be *disjoint* if $\mathcal{A} \cap \mathcal{B} = \emptyset$.
- *Union* The union of two sets $\mathcal{A}$ and $\mathcal{B}$ is the set with elements belonging to $\mathcal{A}$ or $\mathcal{B}$, and is denoted $\mathcal{A} \cup \mathcal{B}$.
- *Set difference and complement* The set difference of $\mathcal{B}$ minus $\mathcal{A}$ is the set with elements those in $\mathcal{B}$ that are not in $\mathcal{A}$, and is denoted $\mathcal{B} \setminus \mathcal{A}$. It is sometimes called the complement of $\mathcal{A}$ in $\mathcal{B}$. The complement of $\mathcal{A}$ in the whole set Ω is often denoted $\mathcal{A}^\mathsf{c}$.
- *Symmetric set difference* The symmetric set difference of $\mathcal{A}$ and $\mathcal{B}$ is defined as the set with elements either in $\mathcal{A}$ or in $\mathcal{B}$, but not in both, and is denoted $\mathcal{A} \triangle \mathcal{B}$.

Sets and set operations can be visualized using a *Venn diagram*. See Figure 1.1 for an example.

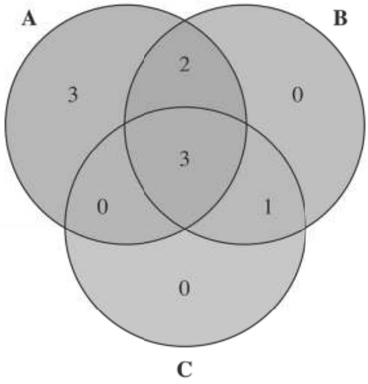

Figure 1.1 A Venn diagram helping visualize the sets $\mathcal{A} = \{1, 2, 4, 5, 6, 7, 8, 9\}$, $\mathcal{B} = \{2, 3, 4, 5, 7, 9\}$, and $\mathcal{C} = \{3, 4, 5, 9\}$. The numbers shown in the figure represent the size of each subset. For example, the intersection of these three sets contains 3 elements, since $\mathcal{A} \cap \mathcal{B} \cap \mathcal{C} = \{4, 5, 9\}$.

Problem 1.2 Prove that $\mathcal{A} \cap \emptyset = \emptyset$, $\mathcal{A} \cup \emptyset = \mathcal{A}$, and $\mathcal{A} \setminus \emptyset = \mathcal{A}$. What is $\mathcal{A} \triangle \emptyset$?

Problem 1.3 Prove that the complement is an involution, i.e., $(\mathcal{A}^\mathsf{c})^\mathsf{c} = \mathcal{A}$.

Problem 1.4 Show that the set difference operation is not symmetric in the sense that $\mathcal{B} \setminus \mathcal{A} \neq \mathcal{A} \setminus \mathcal{B}$ in general. In fact, prove that $\mathcal{B} \setminus \mathcal{A} = \mathcal{A} \setminus \mathcal{B}$ if and only if $\mathcal{A} = \mathcal{B} = \varnothing$.

Proposition 1.5. *The following are true:*

(i) *The intersection operation is commutative, meaning* $\mathcal{A} \cap \mathcal{B} = \mathcal{B} \cap \mathcal{A}$, *and associative, meaning* $(\mathcal{A} \cap \mathcal{B}) \cap \mathcal{C} = \mathcal{A} \cap (\mathcal{B} \cap \mathcal{C})$. *The same is true for the union operation.*
(ii) *The intersection operation is distributive over the union operation, meaning* $(\mathcal{A} \cup \mathcal{B}) \cap \mathcal{C} = (\mathcal{A} \cap \mathcal{C}) \cup (\mathcal{B} \cap \mathcal{C})$.
(iii) *It holds that* $(\mathcal{A} \cap \mathcal{B})^c = \mathcal{A}^c \cup \mathcal{B}^c$. *More generally,* $\mathcal{C} \setminus (\mathcal{A} \cap \mathcal{B}) = (\mathcal{C} \setminus \mathcal{A}) \cup (\mathcal{C} \setminus \mathcal{B})$.

We thus may write $\mathcal{A} \cap \mathcal{B} \cap \mathcal{C}$ and $\mathcal{A} \cup \mathcal{B} \cup \mathcal{C}$, that is, without parentheses, as there is no ambiguity. More generally, for a collection of sets $\{\mathcal{A}_i : i \in I\}$, where I is some index set, we can therefore refer to their intersection and union, denoted

$$\text{(intersection)} \quad \bigcap_{i \in I} \mathcal{A}_i, \qquad \text{(union)} \quad \bigcup_{i \in I} \mathcal{A}_i \,.$$

Remark 1.6 For the reader seeing these operations for the first time, it can be useful to think of $\cap$ and $\cup$ in analogy with the product $\times$ and sum $+$ operations on the integers. In that analogy, $\varnothing$ plays the role of 0.

Problem 1.7 Prove Proposition 1.5. In fact, prove the following identities:

$$\Big(\bigcup_{i \in I} \mathcal{A}_i\Big) \cap \mathcal{B} = \bigcup_{i \in I} (\mathcal{A}_i \cap \mathcal{B}),$$

and

$$\Big(\bigcup_{i \in I} \mathcal{A}_i\Big)^c = \bigcap_{i \in I} \mathcal{A}_i^c, \quad \text{as well as} \quad \Big(\bigcap_{i \in I} \mathcal{A}_i\Big)^c = \bigcup_{i \in I} \mathcal{A}_i^c,$$

for any collection of sets $\{\mathcal{A}_i : i \in I\}$ and any set $\mathcal{B}$.

1.2 Outcomes and Events

Having introduced some elements of Set Theory, we use some of these concepts to define a probability experiment and its possible outcomes.

1.2.1 Outcomes and the Sample Space

In the context of an *experiment*, all the possible *outcomes* are gathered in a *sample space*, denoted Ω henceforth. In mathematical terms, the sample space is a set and the outcomes are elements of that set.

Example 1.8 (Flipping a coin) Suppose that we flip a coin three times in sequence. Assuming the coin can only land heads (H) or tails (T), the sample space Ω consists of all possible ordered sequences of length 3, which in lexicographic order can be written as

$$\Omega = \{\text{HHH}, \text{HHT}, \text{HTH}, \text{HTT}, \text{THH}, \text{THT}, \text{TTH}, \text{TTT}\}.$$

Example 1.9 (Drawing from an urn) Suppose that we draw two balls from an urn in sequence. Assume the urn contains red (R), green (G), and (B) blue balls. If the urn contains at least two balls of each color, or if at each trial the ball is returned to the urn, the sample space Ω consists of all possible ordered sequences of length 2, which in the RGB order can be written as

$$\Omega = \{\text{RR}, \text{RG}, \text{RB}, \text{GR}, \text{GG}, \text{GB}, \text{BR}, \text{BG}, \text{BB}\}. \tag{1.1}$$

If the urn (only) contains one red ball, one green ball, and two or more blue balls, and a ball drawn from the urn is not returned to the urn, the number of possible outcomes is reduced and the resulting sample space is now

$$\Omega = \{\text{RG}, \text{RB}, \text{GR}, \text{GB}, \text{BR}, \text{BG}, \text{BB}\}.$$

Problem 1.10 What is the sample space when we flip a coin five times? More generally, can you describe the sample space, in words and/or mathematical language, corresponding to an experiment where the coin is flipped n times? What is the size of that sample space?

Problem 1.11 Consider an experiment that consists in drawing two balls from an urn that contains red, green, blue, and yellow balls. However, yellow balls are ignored, in the sense that if such a ball is drawn then it is discarded. How does that change the sample space compared to Example 1.9?

While in the previous examples the sample space is finite, the following is an example where it is (countably) infinite.

Example 1.12 (Flipping a coin until the first heads) Consider an experiment where we flip a coin repeatedly until it lands heads. The sample space in this case is

$$\Omega = \{\text{H}, \text{TH}, \text{TTH}, \text{TTTH}, \dots\}.$$

Problem 1.13 Describe the sample space when the experiment consists in drawing repeatedly without replacement from an urn with red, green, and blue balls, three of each color, until a blue ball is drawn.

Remark 1.14 A sample space is in fact only required to contain all possible outcomes. For instance, in Example 1.9 we may always take the sample space to be (1.1) even though in the second situation that space contains outcomes that will never arise.

1.2.2 Events

Events are subsets of Ω that are of particular interest. We say that an event *happens* when the experiment results in an outcome that belongs to the event.

Example 1.15 In the context of Example 1.8, consider the event that the second toss results in heads. As a subset of the sample space, this event is defined as

$$\mathcal{E} = \{\text{HHH}, \text{HHT}, \text{THH}, \text{THT}\}.$$

Example 1.16 In the context of Example 1.9, consider the event that the two balls drawn from the urn are of the same color. This event corresponds to the set

$$\mathcal{E} = \{\text{RR}, \text{GG}, \text{BB}\}.$$

Example 1.17 In the context of Example 1.12, the event that the number of total tosses is even corresponds to the set

$$\mathcal{E} = \{\text{TH}, \text{TTTH}, \text{TTTTTH}, \dots\}.$$

Problem 1.18 In the context of Example 1.8, consider the event that at least two tosses result in heads. Describe this event as a set of outcomes.

1.2.3 Collection of Events

Recall that we are interested in particular subsets of the sample space Ω and that we call these ‘events’. Let Σ denote the collection of events. We assume throughout that Σ satisfies the following conditions:

- The entire sample space is an event, meaning

$$\Omega \in \Sigma. \tag{1.2}$$

- The complement of an event is an event, meaning

$$\mathcal{A} \in \Sigma \;\Rightarrow\; \mathcal{A}^{\mathsf{c}} \in \Sigma. \tag{1.3}$$

- A countable union of events is an event, meaning

$$\mathcal{A}_1, \mathcal{A}_2, \dots \in \Sigma \;\Rightarrow\; \bigcup_{i \ge 1} \mathcal{A}_i \in \Sigma. \tag{1.4}$$

A collection of subsets that satisfies these conditions is called a *σ-algebra*.[2]

Problem 1.19 Suppose that Σ is a σ-algebra. Show that $\emptyset \in \Sigma$ and that a countable intersection of subsets of Σ is also in Σ.

From now on, Σ will denote a σ-algebra over Ω unless otherwise specified. (Note that such a σ-algebra always exists: an example is $\{\emptyset, \Omega\}$.) The pair (Ω, Σ) is then called a *measurable space*.

Remark 1.20 (The power set) The *power set* of Ω, often denoted 2^Ω, is the collection of all its subsets. (Problem 1.49 provides a motivation for this name and notation.) The power set is trivially a σ-algebra. In the context of an experiment with a discrete sample space, it is customary to work with the power set as σ-algebra, because this can always be done without loss of generality (Chapter 2). When the sample space is not discrete, the situation is more complex and the σ-algebra needs to be chosen with more care (Section 3.2).

1.3 Probability Axioms

Before observing the result of an experiment, we speak of the probability that an event will happen. The Kolmogorov axioms formalize this assignment of probabilities to events. This has to be done carefully so that the resulting theory is both coherent and useful for modeling randomness.

A *probability distribution* (aka *probability measure*) on (Ω, Σ) is any real-valued function $\mathbb{P}$ defined on Σ satisfying the following properties or axioms:[3]

- *Non-negativity*

$$\mathbb{P}(\mathcal{A}) \ge 0, \quad \forall \mathcal{A} \in \Sigma.$$

- *Unit measure*

$$\mathbb{P}(\Omega) = 1.$$

[2] This refers to the algebra of sets presented in Section 1.1.

[3] Throughout, we will often use ‘distribution’ or ‘measure’ as shorthand for ‘probability distribution’.

- *Additivity on disjoint events* For any discrete collection of disjoint events $\{\mathcal{A}_i : i \in I\}$,

$$\mathbb{P}\Big(\bigcup_{i \in I} \mathcal{A}_i\Big) = \sum_{i \in I} \mathbb{P}(\mathcal{A}_i). \tag{1.5}$$

A triplet $(\Omega, \Sigma, \mathbb{P})$ with Ω a sample space (a set), Σ a σ-algebra over Ω, and $\mathbb{P}$ a distribution on Σ, is called a *probability space*. We consider such a triplet in what follows.

Problem 1.21 Show that $\mathbb{P}(\varnothing) = 0$ and that

$$0 \le \mathbb{P}(\mathcal{A}) \le 1, \quad \mathcal{A} \in \Sigma.$$

Thus, although nominally a probability distribution takes values in $\mathbb{R}_+$, in fact it takes values in $[0, 1]$.

Proposition 1.22 (Law of Total Probability)**.** *For any two events $\mathcal{A}$ and $\mathcal{B}$,*

$$\mathbb{P}(\mathcal{A}) = \mathbb{P}(\mathcal{A} \cap \mathcal{B}) + \mathbb{P}(\mathcal{A} \cap \mathcal{B}^c). \tag{1.6}$$

Problem 1.23 Prove Proposition 1.22 using the 3rd axiom.

The 3rd axiom applies to events that are disjoint. The following is a corollary that applies more generally. (In turn, this result implies the 3rd axiom.)

Proposition 1.24 (Law of Addition)**.** *For any two events $\mathcal{A}$ and $\mathcal{B}$, not necessarily disjoint,*

$$\mathbb{P}(\mathcal{A} \cup \mathcal{B}) = \mathbb{P}(\mathcal{A}) + \mathbb{P}(\mathcal{B}) - \mathbb{P}(\mathcal{A} \cap \mathcal{B}). \tag{1.7}$$

In particular,

$$\mathbb{P}(\mathcal{A}^c) = 1 - \mathbb{P}(\mathcal{A}), \tag{1.8}$$

and,

$$\mathcal{A} \subset \mathcal{B} \;\Rightarrow\; \mathbb{P}(\mathcal{B} \setminus \mathcal{A}) = \mathbb{P}(\mathcal{B}) - \mathbb{P}(\mathcal{A}). \tag{1.9}$$

Proof We first observe that we can get (1.9) from the fact that $\mathcal{B}$ is the disjoint union of $\mathcal{A}$ and $\mathcal{B} \setminus \mathcal{A}$ and an application of the 3rd axiom.

We now use this to prove (1.7). We start from the disjoint union

$$\mathcal{A} \cup \mathcal{B} = (\mathcal{A} \setminus \mathcal{B}) \cup (\mathcal{B} \setminus \mathcal{A}) \cup (\mathcal{A} \cap \mathcal{B}).$$

Applying the 3rd axiom yields

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

Then $\mathcal{A} \setminus \mathcal{B} = \mathcal{A} \setminus (\mathcal{A} \cap \mathcal{B})$, and applying (1.9), we get

$$\mathbb{P}(\mathcal{A} \setminus \mathcal{B}) = \mathbb{P}(\mathcal{A}) - \mathbb{P}(\mathcal{A} \cap \mathcal{B}),$$

and exchanging the roles of $\mathcal{A}$ and $\mathcal{B}$,

$$\mathbb{P}(\mathcal{B} \setminus \mathcal{A}) = \mathbb{P}(\mathcal{B}) - \mathbb{P}(\mathcal{A} \cap \mathcal{B}).$$

After some cancellations, we obtain (1.7), which then immediately implies (1.8). □

Problem 1.25 (Uniform distribution) Suppose that Ω is finite. For $\mathcal{A} \subset \Omega$, define $\mathbb{U}(\mathcal{A}) = |\mathcal{A}|/|\Omega|$, where $|\mathcal{A}|$ denotes the number of elements in $\mathcal{A}$. Show that $\mathbb{U}$ is a probability distribution on Ω (equipped with its power set, as usual).

1.4 Inclusion-Exclusion Formula

The inclusion-exclusion formula is an expression for the probability of a discrete union of events. We start with some basic inequalities that are directly related to the inclusion-exclusion formula and useful on their own.

Boole's Inequality

Also know as the *union bound*, this inequality[4] is arguably one of the simplest, yet also one of the most useful, inequalities of Probability Theory.

Problem 1.26 (Boole's inequality) Prove that for any countable collection of events $\{\mathcal{A}_i : i \in I\}$,

$$\mathbb{P}\Big(\bigcup_{i \in I} \mathcal{A}_i\Big) \leq \sum_{i \in I} \mathbb{P}(\mathcal{A}_i). \tag{1.10}$$

Note that the right-hand side can be larger than 1 or even infinite. [One possibility is to use a recursion on the number of events, together with Proposition 1.24, to prove the result for any finite number of events. Then pass to the limit to obtain the result as stated.]

Bonferroni's Inequalities

These inequalities[5] comprise Boole's inequality. For two events, we saw the Law of Addition (Proposition 1.24), which is an exact expression for the probability of their union. Consider now three events $\mathcal{A}, \mathcal{B}, \mathcal{C}$. Boole's

[4] Named after George Boole (1815–1864).
[5] Named after Carlo Emilio Bonferroni (1892–1960).