Cambridge University Press & Assessment
978-1-107-17891-5 — Foundations of Agnostic Statistics
Peter M. Aronow, Benjamin T. Miller
Excerpt
More Information

1

Probability Theory

Though there be no such thing as chance in the world, our ignorance of the real cause of any event has the same influence on the understanding.

—David Hume

Our book begins by providing a formalization of probability theory, establishing some basic theorems that will allow us to formally describe random generative processes and quantify relevant features of these processes. We believe that it is important for researchers to understand the assumptions embedded in mathematical probability theory before attempting to make statistical claims. In keeping with the agnostic paradigm, we will attempt to draw attention to the relevant intricacies involved, and highlight where mathematical constructions serve as approximations. The foundation of probability theory allows researchers to precisely define the scope of their research questions and to rigorously quantify their uncertainty about the conclusions they draw from their findings. Our approach is somewhat unconventional, in that we focus on describing random variables *before* we consider data, so as to have well-defined inferential targets. This is sometimes referred to as the "population first" approach.[1] This approach will enable us to engage with the more complex topics in Parts II and III with relative ease and rigor.

We begin this chapter with a discussion of how random generative processes assign probabilities to random events. We can then describe probabilities of individual events and how the probability of one event relates to the probability of another. We proceed to consider random variables, which take on real number values determined by the outcomes of random generative processes. We describe several types of functions that characterize the probability *distributions* of random variables; these distribution functions allow us to characterize the probability that the random variable takes on any given value

[1] See Angrist and Pischke (2009).

or values. When we have two or more random variables whose values are determined simultaneously, we can describe their joint distributions, which also allows us to describe their marginal and conditional distributions. Here, we primarily focus on bivariate relationships (between two random variables), but we outline multivariate generalizations. At the end of the chapter, we provide suggestions for readers interested in a more technical treatment of the material or alternative approaches.

1.1 RANDOM EVENTS

Probability theory is a *mathematical construct* used to represent processes involving randomness, unpredictability, or intrinsic uncertainty. In a setting in which there are several possible outcomes, each with some probability of occurring, we refer to the process by which the outcome is determined as a *random generative process*. In this section, we present the basic principles of probability theory used to describe random generative processes.

1.1.1 What Is Probability?

We can think of a random generative process as a mechanism that selects an *outcome* from among multiple possible outcomes. This mechanism could be flipping a coin or rolling a die, drawing a ball from an urn, selecting a person at random from a group of people, or any other process in which the outcome is in some sense uncertain. A single instance of selecting an outcome is known as a *draw* from or *realization* of the random generative process. The term *experiment* is also commonly used, but we shall refrain from this usage to avoid confusion with experiments in the ordinary sense of the term.

The probability of an event describes the proportion of times that event can be expected to occur among many realizations of a random generative process. This interpretation of probability is known as *frequentist probability* or *frequentism*: the probability of an event A is interpreted as representing how frequently A would occur among many, many draws from a random generative process. It is the long-run average or limiting value of the frequency of observing event A among repeated realizations of the generative process.[2]

Probability theory is a *model*, which is an approximation of reality. Everyday macrophysical processes are not actually characterized by fundamental randomness. Consider, for example, a coin flip. In principle, if we could know the exact mass, shape, and position of the coin at the moment it was flipped and the exact magnitude and direction of the force imparted to it by the flipper

[2] There are other interpretations of probability, most notably the *Bayesian* interpretation, which treats probability as representing a degree of belief or confidence in a proposition. We do not discuss these alternative interpretations and will be operating under the frequentist paradigm throughout this book.

and of all other forces acting on it, we could predict *with certainty* whether it would land on heads or tails.[3]

The mathematical construct of randomness is, therefore, a *modeling assumption*, not necessarily a fundamental feature of reality.[4] It allows us to model the outcomes of the coin flip given our uncertainty about the exact nature of the forces that will act on the coin in any particular instance. Similarly, in the social and health sciences, the assumption of randomness allows us to model various outcomes that we might care about, given our uncertainty about the precise features of complex social or biological processes.

1.1.2 Fundamentals of Probability Theory

We now introduce the formal definitions and notation used to represent the basic elements of random generative processes. There are three formal components that together fully describe a random generative process. The first is the *sample space*, denoted by Ω. The sample space is the set of all possible[5] outcomes of the random generative process. Individual outcomes (sometimes known as *sample points*) are denoted by $\omega \in \Omega$. Outcomes can be represented by numbers, letters, words, or other symbols—whatever is most convenient for describing every distinct possible outcome of the random generative process. For example, if we wanted to describe a single roll of a six-sided die, we could let $\Omega = \{1,2,3,4,5,6\}$. To describe a roll of two six-sided dice, we could let Ω be the set of all ordered pairs of integers between 1 and 6, that is, $\Omega = \{(x,y) \in \mathbb{Z}^2 : 1 \le x \le 6, 1 \le y \le 6\}$. To describe a fair coin flip, we could let $\Omega = \{\text{Heads}, \text{Tails}\}$ or $\Omega = \{H, T\}$. To describe choosing a random person in the United States and measuring their height in inches, we could let Ω be the set of all positive real numbers, $\Omega = \mathbb{R}^+$.

The second component of a random generative process is the *event space*. Events are subsets of Ω and are denoted by capital Roman letters, for example, $A \subseteq \Omega$. Whereas Ω describes all distinguishable states of the world that could result from the generative process, an event may occur in multiple states of the world, so we represent it as a set containing all states of the world in which it occurs. For example, in the case of rolling a single six-sided die, we could represent the event of rolling an even number by the set $A = \{\omega \in \Omega : \omega \text{ is even}\} = \{2,4,6\}$. Of course, an event can also correspond to a single state of the world, for instance, the event of rolling a 3, which we might represent by

[3] See Diaconis, Holmes, and Montgomery (2007).

[4] Current thinking in physics suggests that randomness *is* a fundamental feature of quantum-mechanical processes rather than merely a representation of unknown underlying variables determining individual outcomes. We are not considering quantum mechanics in this book. Suffice it to say that quantum randomness is probably not relevant to the social or health sciences.

[5] We use the word "possible" here loosely, as the probability of a given outcome occurring may be zero.

the set $B = \{3\}$; such events are variously known as *atomic events* , *elementary events*, or *simple events*. For a given random generative process, a set of events is called an event space if it satisfies certain properties, which we state in the following definition. (We use the notation A^C to refer to the *complement* of the event A with respect to the sample space: $A^C = \Omega \backslash A = \{\omega \in \Omega : \omega \notin A\}$.)

Definition 1.1.1. *Event Space*
A set S of subsets of Ω is an *event space* if it satisfies the following:

- Nonempty: $S \neq \emptyset$.
- Closed under complements: if $A \in S$, then $A^C \in S$.
- Closed under countable unions: if $A_1, A_2, A_3, ... \in S$, then $A_1 \cup A_2 \cup A_3 \cup ... \in S$.[6]

Since each event in an event space will have an associated probability of occurring, these properties ensure that certain types of events will always have well-defined probabilities. Consider, again, a single roll of a six-sided die. Suppose A is the event of rolling an even number. If we can assign a probability to this event occurring, then we will also be able to assign a probability to this event *not* occurring, that is, A^C, the event of rolling an odd number. Similarly, suppose B is the event of rolling a number greater than 4. If we can assign probabilities to A and B, then we will also be able to assign a probability to the event of *at least one* of these occurring, that is, $A \cup B$, the event of rolling a 2, 4, 5, or 6.

This brings us to the final component needed to mathematically describe a random generative process: the *probability measure*. A probability measure is a function $P : S \to \mathbb{R}$ that assigns a probability to every event in the event space.[7] To ensure that P assigns probabilities to events in a manner that is coherent and in accord with basic intuitions about probabilities, we must place some conditions on P. Such conditions are provided by the *Kolmogorov probability axioms*, which serve as the foundation of probability theory. These axioms define a *probability space*, a construct that both accords with basic intuitions about probabilities and lends itself to rigorous and useful mathematics.

[6] A set S of subsets of another set Ω that satisfies these properties is formally known as a σ-*algebra* or σ-*field* on Ω. Some readers well versed in set theory may wonder why we do not simply let $S = \mathcal{P}(\Omega)$, the power set (that is, the set of all subsets) of Ω. For reasons that we will not discuss in this book, this does not always work; for some sample spaces Ω, it is impossible to define the probability of every subset of Ω in a manner consistent with the axioms of probability (see Definition 1.1.2). We need not worry too much about this point, though; in practice we will be able to define the probability of any event of interest without much difficulty. For this reason, we suggest that readers not worry too much about σ-algebras on their first read.

[7] Note: we do not make the stronger assumption that $P : S \to [0, 1]$, since we prove this in Theorem 1.1.4.

Definition 1.1.2. *Kolmogorov Axioms*
Let Ω be a sample space, S be an event space, and P be a probability measure. Then (Ω, S, P) is a *probability space* if it satisfies the following:

- Non-negativity: $\forall A \in S$, $P(A) \geq 0$, where $P(A)$ is finite and real.
- Unitarity: $P(\Omega) = 1$.
- Countable additivity: if $A_1, A_2, A_3, ... \in S$ are pairwise disjoint,[8] then

$$P(A_1 \cup A_2 \cup A_3 \cup ...) = P(A_1) + P(A_2) + P(A_3) + ... = \sum_i P(A_i).$$

The intuition behind each of these axioms is as follows: The first axiom states that the probability of any event is a non-negative number; there cannot be a less-than-zero chance of an event occurring. The second axiom states that the probability measure of the entire sample space is one.[9] In other words, it is certain that *some* outcome will occur. Finally, the third axiom states that, given any number of *mutually exclusive* events, the probability that one of those events will occur is the sum of their individual probabilities. Together, these axioms are sufficient to rule out any probability statements that would be nonsensical, and they provide the building blocks that will allow us to derive other useful properties (as we will see in Theorem 1.1.4, *Basic Properties of Probability*).

We can represent any random generative process as a probability space (Ω, S, P), as illustrated by the following simple example.

Example 1.1.3. *A Fair Coin Flip*
Consider a fair coin flip. Let H represent the outcome "heads" and T represent the outcome "tails." Let $\Omega = \{H, T\}$ and $S = \{\emptyset, \{H\}, \{T\}, \{H, T\}\}$. Then we can define the probability measure as follows:

$$P(A) = \frac{1}{2}|A|, \forall A \in S.$$

The notation $|A|$ denotes the *cardinality* of the set A, that is, the number of elements in A. So this means

- $P(\emptyset) = \frac{1}{2}|\emptyset| = \frac{1}{2} \cdot 0 = 0$. The probability of nothing happening is zero.
- $P\big(\{H\}\big) = \frac{1}{2}\big|\{H\}\big| = \frac{1}{2} \cdot 1 = \frac{1}{2}$. The probability of getting heads is $\frac{1}{2}$.
- $P\big(\{T\}\big) = \frac{1}{2}\big|\{T\}\big| = \frac{1}{2} \cdot 1 = \frac{1}{2}$. The probability of getting tails is $\frac{1}{2}$.

[8] Recall that sets A and B are disjoint if $A \cap B = \emptyset$. We say that $A_1, A_2, A_3, ...$ are pairwise disjoint if each of them is disjoint from every other, that is, $\forall i \neq j$, $A_i \cap A_j = \emptyset$.

[9] Notice that Definition 1.1.1 implies that any event space must contain Ω: S is nonempty, so $\exists A \in S$. Since S is closed under complements, $A^C \in S$, and so since S is closed under countable unions, $A \cup A^C = \Omega \in S$. Likewise, $\Omega^C = \emptyset \in S$.

- $P\big(\{H,T\}\big) = \frac{1}{2}\big|\{H,T\}\big| = \frac{1}{2}\cdot 2 = 1$. The probability of getting either heads or tails is one.

The reader can verify that S is a proper event space (that is, it is nonempty and closed under complements and countable unions) and that P satisfies the Kolmogorov axioms, so (Ω, S, P) is a probability space. △[10]

Several other fundamental properties of probability follow directly from the Kolmogorov axioms.

Theorem 1.1.4. *Basic Properties of Probability*
Let (Ω, S, P) be a probability space.[11] Then

- Monotonicity: $\forall A, B \in S$, if $A \subseteq B$, then $P(A) \leq P(B)$.
- Subtraction rule: $\forall A, B \in S$, if $A \subseteq B$, then $P(B \backslash A) = P(B) - P(A)$.
- Zero probability of the empty set: $P(\emptyset) = 0$.
- Probability bounds: $\forall A \in S$, $0 \leq P(A) \leq 1$.
- Complement rule: $\forall A \in S$, $P\big(A^C\big) = 1 - P(A)$.

Proof: Let $A, B \in S$ with $A \subseteq B$. Since $B = A \cup (B \backslash A)$, and A and $(B \backslash A)$ are disjoint, countable additivity implies

$$P(B) = P(A) + P(B \backslash A).$$

Rearranging this equation, non-negative probabilities then imply monotonicity: $P(B \backslash A) \geq 0$, so

$$P(A) = P(B) - P(B \backslash A) \leq P(B).$$

Rearranging again yields the subtraction rule:

$$P(B \backslash A) = P(B) - P(A).$$

The subtraction rule, in turn, implies zero probability of the empty set: $A \subseteq A$, so

$$P(\emptyset) = P(A \backslash A) = P(A) - P(A) = 0.$$

Monotonicity and unitarity (and non-negativity) imply the probability bounds: since $A \subseteq \Omega$,

$$0 \leq P(A) \leq P(\Omega) = 1.$$

Finally, the subtraction rule and unitarity imply the complement rule:

$$P\big(A^C\big) = P(\Omega \backslash A) = P(\Omega) - P(A) = 1 - P(A). \quad \square$$

[10] Note that we use the △ symbol to denote the end of an example.

[11] This assumption shall henceforth be implicit in all definitions and theorems referring to Ω, *S*, and/or *P*.

We can put each of these properties in simple terms. Monotonicity implies that, if one event is a subset of another (so that the former always occurs whenever the latter does), then the probability of the former occurring is no greater than that of the latter. The subtraction rule implies that the probability that the second event occurs but not the first is equal to the probability of the second event minus the probability of the first event. Zero probability of the empty set means that *some* event in our event space must occur, and probability bounds mean that each of these events has some probability of occurring between zero and one. Finally, the complement rule implies that the probability of any of these events *not* occurring is one minus the probability of the event occurring—so that the probability that a given event either occurs or does not occur is one.

1.1.3 Joint and Conditional Probabilities

We often want to describe how the probability of one event relates to the probability of another. We begin by establishing the *joint probability* of events A and B, or the probability that events A and B will both occur in a single draw from (Ω, S, P).

Definition 1.1.5. *Joint Probability*
For $A, B \in S$, the *joint probability* of A and B is $P(A \cap B)$.

In other words, the joint probability of two events A and B is the probability of the intersection of A and B (which is itself an event in S), that is, the set of all states of the world in which both A and B occur. We illustrate this point with the following example.

Example 1.1.6. *A Fair Die Roll*
Consider a roll of one fair (six-sided) die. Let $\Omega = \{1,2,3,4,5,6\}$, $S = \mathcal{P}(\Omega)$ (the power set—that is, the set of all subsets—of Ω), and $P(A) = \frac{1}{6}|A|$, $\forall A \in S$. Let $A = \{\omega \in \Omega : \omega \geq 4\} = \{4,5,6\}$ and $B = \{\omega \in \Omega : \omega \text{ is even}\} = \{2,4,6\}$. Then

$$P(A \cap B) = P\big(\{4,5,6\} \cap \{2,4,6\}\big) = P\big(\{4,6\}\big) = \frac{1}{6}\big|\{4,6\}\big| = \frac{2}{6} = \frac{1}{3}. \ \triangle$$

Just as $P(A \cap B)$ is the probability that both A *and* B will occur in a single draw from (Ω, S, P), $P(A \cup B)$ is the probability that A *or* B (or both) will occur in a single draw from (Ω, S, P). The following theorem allows us to relate these two probabilities.

Theorem 1.1.7. *Addition Rule*
For $A, B \in S$,

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

Proof: Note that $(A\backslash B)$, $(B\backslash A)$, and $(A\cap B)$ are pairwise disjoint and

$$(A\cup B)=(A\backslash B)\cup(B\backslash A)\cup(A\cap B),$$

so by countable additivity,

$$\begin{aligned}P(A\cup B)&=P(A\backslash B)+P(B\backslash A)+P(A\cap B)\\&=P\big(A\backslash(A\cap B)\big)+P\big(B\backslash(A\cap B)\big)+P(A\cap B)\\&=P(A)-P(A\cap B)+P(B)-P(A\cap B)+P(A\cap B)\\&=P(A)+P(B)-P(A\cap B),\end{aligned}$$

where the second equality holds because

$$\begin{aligned}A\backslash B&=\emptyset\cup(A\backslash B)\\&=\left(A\cap A^C\right)\cup\left(A\cap B^C\right)\\&=A\cap\left(A^C\cup B^C\right)\\&=A\cap(A\cap B)^C\\&=A\backslash(A\cap B),\end{aligned}$$

and likewise $B\backslash A=B\backslash(A\cap B)$, while the third equality follows from the subtraction rule, since $A\cap B\subseteq A$ and $A\cap B\subseteq B$. □

In other words, the probability of *at least one* of two events occurring is equal to the sum of the probabilities of *each* occurring minus the probability of *both* occurring. Of course, if A and B are disjoint, this reduces to $P(A\cup B)=P(A)+P(B)$, which is just a special case of countable additivity.

We might also want to describe the probability of observing event A *given* that we observe event B. This is known as *conditional probability*.

Definition 1.1.8. *Conditional Probability*
For $A,B\in S$ with $P(B)>0$, the *conditional probability* of A given B is

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

We can rearrange this definition to obtain another useful formula: the *Multiplicative Law of Probability*.

Theorem 1.1.9. *Multiplicative Law of Probability*
For $A,B\in S$ with $P(B)>0$,

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

Proof: Rearrange Definition 1.1.8. □

One of the most important theorems regarding conditional probability is *Bayes' Rule* (also known as *Bayes' Theorem* or *Bayes' Law*). Bayes' Rule relates the conditional probability of A given B to the conditional probability of B given A. Suppose we have a hypothesis about the probability that some event A will occur, and we then observe event B. We can update the probability that we predict A will occur, using information about the frequency with which B occurs given A. This kind of deduction constitutes a key element of probabilistic reasoning and thus has many applications in the social sciences.

Theorem 1.1.10. *Bayes' Rule*
For $A, B \in S$ with $P(A) > 0$ and $P(B) > 0$,

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

Proof: By the Multiplicative Law of Probability, $P(A \cap B) = P(B|A)P(A)$. So, by the definition of conditional probability,

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}. \quad \square$$

Although we will not be making use of Bayes' Rule in this book, we would be remiss to neglect it entirely. The following example illustrates the above concepts, including the utility of Bayes' Rule, by combining the previous examples of fair coins and fair dice.

Example 1.1.11. *Flipping a Coin and Rolling a Die*
Consider the following generative process. An experimenter flips a fair coin. If the coin comes up heads, the experimenter rolls a fair four-sided die. If the coin comes up tails, the experimenter rolls a fair six-sided die. The sample space can thus be represented by

$$\Omega = \{(H,1),(H,2),(H,3),(H,4),(T,1),(T,2),(T,3),(T,4),(T,5),(T,6)\}.$$

Let A denote the event of observing heads, B denote the event of observing 3, and C denote the event of observing 6. Formally, $A = \{(H,1),(H,2),(H,3),(H,4)\}$, $B = \{(H,3),(T,3)\}$, and $C = \{(T,6)\}$. What is the (joint) probability of observing heads and 3? The probability of observing heads is $P(A) = \frac{1}{2}$. Additionally, if heads is observed, then the experimenter rolls a fair four-sided die, so the probability of observing 3 *given that heads has been observed* is $P(B|A) = \frac{1}{4}$. So, by the Multiplicative Law of Probability,

$$P(A \cap B) = P(B|A)P(A) = \frac{1}{4} \cdot \frac{1}{2} = \frac{1}{8}.$$

Likewise, the probability of observing tails and 3 is

$$P(A^C \cap B) = P\left(B \,\middle|\, A^C\right)P\left(A^C\right) = \frac{1}{6} \cdot \frac{1}{2} = \frac{1}{12}.$$

Cambridge University Press & Assessment
978-1-107-17891-5 — Foundations of Agnostic Statistics
Peter M. Aronow, Benjamin T. Miller
Excerpt
More Information

Since the experimenter can never roll a 6, if the coin comes up heads, the probability of observing heads and 6 is

$$P(A \cap C) = P(\emptyset) = 0.$$

The conditional probability of observing 3 given that heads (or tails) was observed is relatively straightforward, as we see above. But suppose we wanted to know the conditional probability that heads was observed given that 3 is observed. This is where Bayes' Rule is useful. We want to know $P(A|B)$. We know $P(B|A)$ and $P(A)$. What is $P(B)$? From countable additivity,

$$P(B) = P(A \cap B) + P\left(A^C \cap B\right) = \frac{1}{8} + \frac{1}{12} = \frac{5}{24}.$$

Thus, by Bayes' Rule,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{\frac{1}{4} \cdot \frac{1}{2}}{\frac{5}{24}} = \frac{3}{5}. \triangle$$

The "trick" used above to calculate $P(B)$ is actually a special case of another important theorem, the *Law of Total Probability*. To state this theorem, we require the following definition.

Definition 1.1.12. *Partition*
If $A_1, A_2, A_3, ... \in S$ are nonempty and pairwise disjoint, and $\Omega = A_1 \cup A_2 \cup A_3 \cup ...$, then $\{A_1, A_2, A_3, ...\}$ is a *partition* of Ω.

A partition divides the sample space into mutually exclusive and exhaustive categories or "bins."[12] Every outcome in Ω is contained in exactly one A_i, so exactly one event A_i in the partition occurs for any draw from (Ω, S, P).

Theorem 1.1.13. *Law of Total Probability*
If $\{A_1, A_2, A_3, ...\}$ is a partition of Ω and $B \in S$, then

$$P(B) = \sum_i P(B \cap A_i).$$

If we also have $P(A_i) > 0$ for $i = 1, 2, 3, ...$, then this can also be stated as

$$P(B) = \sum_i P(B|A_i)P(A_i).$$

Proof: If $\{A_1, A_2, A_3, ...\}$ is a partition of Ω, then $\forall i \neq j$,

$$(B \cap A_i) \cap (B \cap A_j) = (B \cap B) \cap (A_i \cap A_j) = B \cap (A_i \cap A_j) = B \cap \emptyset = \emptyset.$$

[12] The number of bins may be finite or countably infinite.