1

Foundations of Probability Theory

For centuries, mankind lived with the idea that uncertainty was the domain of the gods and fell beyond the reach of human calculation. Common gambling led to the first steps toward understanding probability theory, and the colorful Italian mathematician and physician Gerolamo Cardano (1501–1575) was the first to attempt a systematic study of the calculus of probabilities. As an ardent gambler himself, Cardano wrote a handbook for fellow gamblers entitled *Liber de Ludo Aleae* (The Book of Games of Chance) about probabilities in games of chance like dice. He originated and introduced the concept of the set of outcomes of an experiment, and for cases in which all outcomes are equally probable, he defined the probability of any one event occurring as the ratio of the number of favorable outcomes to the total number of possible outcomes. This may seem obvious today, but in Cardano's day such an approach marked an enormous leap forward in the development of probability theory.

Nevertheless, many historians mark 1654 as the birth of the study of probability, since in that year questions posed by gamblers led to an exchange of letters between the great French mathematicians Pierre de Fermat (1601–1665) and Blaise Pascal (1623–1662). This famous correspondence laid the groundwork for the birth of the study of probability, especially their question of how two players in a game of chance should divide the stakes if the game ends prematurely. This problem of points, which will be discussed in Chapter 3, was the catalyst that enabled probability theory to develop beyond mere combinatorial enumeration.

In 1657, the Dutch astronomer Christiaan Huygens (1629–1695) learned of the Fermat–Pascal correspondence and shortly thereafter published the book *De Ratiociniis de Ludo Aleae* (On Reasoning in Games of Chance), in which he worked out the concept of expected value and unified various problems that had been solved earlier by Fermat and Pascal. Huygens' work led the field for many years until, in 1713, the Swiss mathematician Jakob Bernoulli

(1654–1705) published *Ars Conjectandi* (The Art of Conjecturing) in which he presented the first general theory for calculating probabilities. Then, in 1812, the great French mathematician Pierre Simon Laplace (1749–1827) published his *Théorie Analytique des Probabilités*. Here, Laplace applied probabilistic ideas to many scientific and practical problems, and his book represents perhaps the single greatest contribution in the history of probability theory. Towards the end of the nineteenth century, mathematicians attempted to construct a solid foundation for the mathematical theory of probability by defining probabilities in terms of relative frequencies. That attempt, however, was marked by much controversy, and the frequency view of probability did not lead to a satisfactory theory.

An acceptable definition of probability should be precise enough for use in mathematics and yet comprehensive enough to be applicable to a wide range of phenomena. It was not until 1933 that the great Russian mathematician Andrey Nikolaevich Kolmogorov (1903–1987) laid a satisfactory mathematical foundation for probability theory by taking a number of axioms as his starting point, as had been done in other fields of mathematics. Axioms state a number of minimal requirements that the mathematical objects in question (such as points and lines in geometry) must satisfy. In the axiomatic approach of Kolmogorov, probability figures as a function on subsets of a so-called sample space, where the sample space represents the set of all possible outcomes of the experiment. He assigned probabilities to these subsets in a consistent way, using the concept of additivity from measure theory. Those axioms are the basis for the mathematical theory of probability, and, as a milestone, are sufficient to logically deduce the law of large numbers. This law confirms our intuition that the probability of an event in a repeatable experiment can be estimated by the relative frequency of its occurrence in many repetitions of the experiment. The law of large numbers is the fundamental link between theory and the real world.

Nowadays, more than ever before, probability theory is indispensable in a wide variety of fields. It is absolutely essential to the field of insurance. Likewise the stock market, "the largest casino in the world," cannot do without it. Call centers and airline companies apply probabilistic methods to determine how many service desks will be needed based on expected demand. In stock control, probability theory is used to find a balance between the cost risks of running out of inventory and the cost of holding too much inventory in a market with uncertain demand. Engineers use probability theory when constructing dikes, to calculate the probability of water levels exceeding their margins. Judges and physicians benefit from a basic knowledge of probability theory, to help make better judicial and medical decisions. In short, probabilistic thinking has become an integral part of modern life.

The purpose of this chapter is to give the student a solid basis for solving probability problems. We first discuss the intuitive and fundamental axioms of probability, and from them derive a number of basic rules for the calculation of probabilities. These rules include the complement rule, the addition rule, and the inclusion–exclusion rule, for which many illustrative examples are provided. These examples, which are instructive and provide insight into the theory, include classical probability problems such as the birthday problem and the hat-check problem.

Traditionally, introductory probability books begin with a comprehensive discussion of set theory and combinatorics before presenting the "real stuff." This will not be done in this book, as it is not necessary and often stifles the student's natural curiosity and fascination with probability. Rather, the appendices at the end of the book provide a self-contained overview of the essentials of set theory and the basic tools from combinatorics, and these tools are not introduced until they are needed.

1.1 Probabilistic Foundations

A probability model is a mathematical representation of a real-world situation or a random experiment. It consists of a complete description of all possible outcomes of the experiment and an assignment of probabilities to these outcomes. The set of all possible outcomes of the experiment is called the *sample space*. A sample space is always such that one and only one of the possible outcomes occurs if the experiment is performed. Here are some examples:

- The experiment is to toss a coin once. The sample space is the set $\{H, T\}$, where H means that the outcome of the toss is a head and T that it is a tail. Each outcome gets assigned a probability of $\frac{1}{2}$ if the coin is fair.
- The experiment is to roll a die once. The sample space is the set $\{1, 2, \ldots, 6\}$, where outcome i means that i dots appear on the up face. Each outcome gets assigned a probability of $\frac{1}{6}$ if the die is fair.
- The experiment is to choose a letter at random from the word "statistics." The sample space is the set $\{s, t, a, i, c\}$. The probabilities $\frac{3}{10}$, $\frac{3}{10}$, $\frac{1}{10}$, $\frac{2}{10}$, and $\frac{1}{10}$ are assigned to the five outcomes s, t, a, i, and c.
- The experiment is to repeatedly roll a fair die until the first six shows up. The sample space is the set $\{1, 2, \ldots\}$ of the positive integers. Outcome k indicates that the first six shows up on the kth roll. The probabilities $\frac{1}{6}, \frac{5}{6} \times \frac{1}{6}, (\frac{5}{6})^2 \times \frac{1}{6}, \ldots$ are assigned to the outcomes $1, 2, 3, \ldots$.

- The experiment is to measure the time until the first emission of a particle from a radioactive source. The sample space is the set $(0, \infty)$ of the positive real numbers, where outcome t indicates that it takes a time t until the first emission of a particle. Taking an appropriate unit of time, the probability $\int_a^b e^{-t}\,dt$ can be assigned to each time interval (a, b) on the basis of physical properties of radioactive material, where $e = 2.71828\ldots$ is the base of the natural logarithm.

In probability applications we are typically interested in particular subsets of the sample space, which in probability language are called *events*. The terms event and subset of outcomes of an experiment are used interchangeably in probability theory. In the second example, the event that an odd number is rolled is the subset $A = \{1, 3, 5\}$ of the sample space. In the fourth example, the event that more than six rolls are needed to get a six is the subset $A = \{7, 8, \ldots\}$ of the sample space. In the fifth example, the event that it takes between 5 and 7 time units until the first emission of a particle is the subset $A = \{t : \ 5 \leq t \leq 7\}$ of the sample space, where ":" means "such that."

Various choices for the sample space are sometimes possible. In the experiment of tossing a fair coin twice, a possible choice for the sample space is the set $\{HH, HT, TH, TT\}$. Another possible choice is the set $\{0, 1, 2\}$, where the outcome indicates the number of heads obtained. The assignment of probabilities to the elements of the sample space differs for the two choices. In the first choice of the sample space, the four elements are equally likely and each element gets assigned the probability $\frac{1}{4}$. In the second choice of the sample space, the elements are not equally likely and the elements 0, 1, and 2 get assigned the probabilities $\frac{1}{4}$, $\frac{1}{2}$, and $\frac{1}{4}$. In general, it is preferable to use a sample space with equally likely outcomes whenever possible.

In the first three examples above, the sample space is a finite set. In the fourth example the sample space is a so-called countably infinite set, while in the fifth example the sample space is a so-called uncountable set. Let us briefly explain these basic concepts from set theory, see also Appendix B. The set of natural numbers (positive integers) is an infinite set and is the prototype of a countably infinite set. In general, a nonfinite set is called *countably infinite* if a one-to-one function exists which maps the elements of the set to the set of natural numbers. In other words, every element of the set can be assigned to a unique natural number and conversely each natural number corresponds to a unique element of the set. For example, the set of squared numbers $1, 4, 9, 16, 25, \ldots$ is countably infinite. Not all sets with an infinite number of elements are countably infinite. The set of all points on a line and the set of all real numbers between 0 and 1 are examples of infinite sets that are not countable.

The German mathematician Georg Cantor (1845–1918) proved this result in the nineteenth century. This discovery represented an important milestone in the development of mathematics and logic (the concept of infinity, to which even scholars from ancient Greece had devoted considerable energy, obtained a solid theoretical basis for the first time through Cantor's work). Sets that are neither finite nor countably infinite are called *uncountable*, whereas sets that are either finite or countably infinite are called *countable*.

1.1.1 Axioms of Probability Theory

A probability model consists of a sample space together with the assignment of probability, where probability is a function that assigns numbers between 0 and 1 to subsets of the sample space. The axioms of probability are mathematical rules that the probability function must satisfy. The axioms of probability are essentially the same for a chance experiment with a countable or an uncountable sample space. A distinction must be made, however, between the sorts of subsets to which probabilities can be assigned, whether these subsets occur in countable or uncountable sample spaces. In the case of a countable sample space, probabilities can be assigned to each subset of the sample space. In the case of an uncountable sample space, weird subsets can be constructed to which we cannot associate a probability. Then the probability measure is only defined on a sufficiently rich collection of well-behaved subsets, see Appendix B for more details. These technical matters will not be discussed further in this introductory book. The reader is asked to accept the fact that, for more fundamental mathematical reasons, probabilities can only be assigned to well-behaved subsets when the sample space is uncountable. In the case that the uncountable sample space is the set of real numbers, then essentially only those subsets consisting of a finite interval, the complement of any finite interval, the union of any countable number of finite intervals, or the intersection of any countable number of finite intervals are assigned a probability. These subsets suffice for practical purposes. The probability measure on the sample space is denoted by P. It assigns to each well-behaved subset A a probability $P(A)$ and must satisfy the following properties:

Axiom 1 $P(A) \geq 0$ *for each subset* A *of the sample space.*

Axiom 2 $P(\Omega) = 1$ *for the sample space* Ω.

Axiom 3 $P\big(\bigcup_{i=1}^{\infty} A_i\big) = \sum_{i=1}^{\infty} P(A_i)$ *for every collection of pairwise disjoint subsets* $A_1, A_2, \ldots$ *of the sample space.*

The *union* $\bigcup_{i=1}^{\infty} A_i$ of the subsets $A_1, A_2, \ldots$ is defined as the set of all outcomes which belong to at least one of the subsets $A_1, A_2, \ldots$. The subsets $A_1, A_2, \ldots$ are said to be *pairwise disjoint* when any two subsets have no element in common. In probability terms, any subset of the sample space is called an *event*. If the outcome of the chance experiment belongs to the subset A, then the event A is said to occur. The events $A_1, A_2, \ldots$ are said to be *mutually exclusive* (or *disjoint*) if the corresponding sets $A_1, A_2, \ldots$ are pairwise disjoint. In other words, events $A_1, A_2, \ldots$ are mutually exclusive if the occurrence of one of these events implies the non-occurrence of the others. For example, suppose that a die is rolled. The outcome is one of the numbers from 1 to 6. Let A be the event that the outcome 1 or 2 occurs and B be the event that the outcome 5 or 6 occurs. Then A and B are mutually exclusive events. As another example, suppose that a coin is tossed until heads appears for the first time. Let A_k be the event that heads appears for the first time at the kth toss. Then the events $A_1, A_2, \ldots$ are mutually exclusive.

The first two axioms simply express a probability as a number between 0 and 1. The crucial third axiom states that, for any infinite sequence of mutually exclusive events, the probability of at least one of these events occurring is the sum of their individual probabilities. This property also holds for any *finite* sequence of mutually exclusive events. Using the concept of the empty set, the proof of this result is almost trivial, see Rule 1.1 in Section 1.5. The countable additivity in Axiom 3 is required to have a unified framework for finite and nonfinite sample spaces. Starting with the three axioms and a few definitions, a powerful and beautiful theory of probability can be developed.

The standard notation for the sample space is the symbol Ω. An outcome of the sample space is denoted by ω. A sample space together with a collection of events and an assignment of probabilities to the events is called a *probability space*. For a countable sample space Ω, it is sufficient to assign a probability $p(\omega)$ to each element $\omega \in \Omega$ such that $p(\omega) \geq 0$ and $\sum_{\omega \in \Omega} p(\omega) = 1$. A probability measure P on Ω is then defined by specifying the probability of each event A as

$$P(A) = \sum_{\omega \in A} p(\omega).$$

In other words, $P(A)$ is the sum of the individual probabilities of the outcomes ω that belong to the set A. It is left to the reader to verify that P satisfies Axioms 1 to 3.

A probability model is constructed with a specific situation or experiment in mind. The assignment of probabilities is part of the translation process from a concrete context into a mathematical model. Probabilities may be assigned to

events any way you like, as long as the above axioms are satisfied. To make your choice of the probabilities useful, the assignment should result in a "good" model for the real-world situation. There are two main approaches to assigning probabilities to events. In the relative-frequency approach, probabilities are assigned to the outcomes of a physical experiment having the feature that it can be repeated over and over under identical conditions. Think of spinning a roulette wheel or rolling dice. Then one may speak of *physical probabilities* and such probabilities can be determined experimentally. In the subjective approach, the word probability is roughly synonymous with plausibility and probability is defined as the degree of belief a particular person holds in the occurrence of an event. Think of the chances of your favorite horse winning a race or the chances of your favorite baseball team winning the World Series. Hence judgment is used as the basis for assigning *subjective probabilities*. The use of the subjective approach is usually limited to experiments that are unrepeatable. In this book the emphasis is on physical probabilities, but we will also pay attention to subjective probabilities in Chapters 2 and 7.

1.2 Classical Probability Model

In many experiments with finitely many outcomes $\omega_1, \ldots, \omega_m$, it is natural to assume that all these outcomes are equally likely to occur (this probability model is, in fact, also based on judgment). In such a case, $p(\omega_i) = \frac{1}{m}$ for $i = 1, \ldots, m$ and each event A gets assigned the probability

$$P(A) = \frac{m(A)}{m},$$

where $m(A)$ is the number of outcomes in the set A. This model is sometimes called the *classical probability model* or the *Laplace model*.

Example 1.1 John, Pedro, and Rosita each roll one fair die. How do we calculate the probability that Rosita's score is equal to the sum of the scores of John and Pedro?

Solution. We take the set $\{(i,j,k) : i,j,k = 1,\ldots,6\}$ as sample space for the chance experiment, where the outcome (i,j,k) means that John's score is i dots, Pedro's score is j dots, and Rosita's score is k dots. Each of the 216 possible outcomes is equally probable, and thus gets assigned a probability mass of $\frac{1}{216}$. Rosita's score is equal to the sum of the scores of John and Pedro if and only if one of the 15 outcomes (1,1,2), (1,2,3), (2,1,3), (1,3,4), (3,1,4), (2,2,4), (1,4,5), (4,1,5), (2,3,5), (3,2,5), (1,5,6), (5,1,6), (2,4,6), (4,2,6), (3,3,6) occurs. The probability of this event is thus $\frac{15}{216}$.

Example 1.2 Three players enter a room and are given a red or a blue hat to wear. The color of each hat is determined by a fair coin toss. Players cannot see the color of their own hats, but do see the color of the other two players' hats. The game is won when at least one of the players correctly guesses the color of his own hat and no player gives an incorrect answer. In addition to having the opportunity to guess a color, players may also pass. Communication of any kind between the players is not permissible after they have been given their hats; however, they may agree on a group strategy beforehand. The players decided upon the following strategy. A player who sees that the other two players wear a hat with the same color guesses the opposite color for his/her own hat; otherwise, the player says nothing. What is the probability of winning the game under this strategy?

Solution. This chance experiment can be seen as tossing a fair coin three times. As sample space, we take the set consisting of the eight elements *RRR*, *RRB*, *RBR*, *BRR*, *BBB*, *BBR*, *BRB*, *RBB*, where *R* stands for a red hat and *B* for a blue hat. Each element of the sample space is equally probable and gets assigned a probability of $\frac{1}{8}$. The strategy is winning if one of the six outcomes *RRB*, *RBR*, *BRR*, *BBR*, *BRB*, or *RBB* occurs (verify!). Thus, the probability of winning the game under the chosen strategy is $\frac{3}{4}$.

In Example 1.2 we have encountered a useful problem-solving strategy: see whether the problem can be related to a familiar problem.

As preparation for the next example, consider a task that involves a sequence of r choices. Suppose that n_1 is the number of possible ways the first choice can be made, n_2 is the number of possible ways the second choice can be made after the first choice has been made, and n_3 is the number of possible ways the third choice can be made after the first two choices have been made, etc. Then the total number of possible ways the task can be performed is $n_1 \times n_2 \times \cdots \times n_r$. For example, the total number of possible ways five people can stand in line is $5 \times 4 \times 3 \times 2 \times 1 = 120$. In other words, there are 120 permutations.

Example 1.3 In a Monte Carlo casino the roulette wheel is divided into 37 sections numbered 1 to 36 and 0. What is the probability that all numbers showing up in eight spins of the wheel are different?

Solution. Take as sample space the set of all ordered sequences $(i_1, \ldots, i_8)$, where i_k is the number showing up at the kth spin of the wheel. The sample space has $37 \times 37 \times \cdots \times 37 = 37^8$ equiprobable elements. The number of elements for which all components are different is $37 \times 36 \times \cdots \times 30$. Therefore, the sought probability is

$$\frac{37 \times 36 \times \cdots \times 30}{37^8} = 0.4432.$$

We reiterate the concept of the binomial coefficient before continuing.

Definition 1.1 *The binomial coefficient* $\binom{n}{k}$ *denotes the total number of ways to choose* k *different objects out of* n *distinguishable objects, with order not mattering.*

In other words, $\binom{n}{k}$ is the total number of combinations of k different objects out of n. The key difference between permutations and combinations is *order*. Combinations are *unordered* selections, permutations are *ordered* arrangements. In Appendix A these important concepts are discussed extensively and illustrated with several combinatorial probability problems. For any integers n and k with $1 \leq k \leq n$, the binomial coefficient can be calculated as

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!},$$

where $m!$ is shorthand for $1 \times 2 \times \cdots \times m$ with the convention $0! = 1$. Note that $\binom{n}{k} = \binom{n}{n-k}$ with the convention $\binom{n}{0} = 1$. For example, the number of ways to choose three jurors out of five candidates is $\binom{5}{3} = \frac{5!}{3!\,2!} = \frac{120}{6 \times 2} = 10$, with order not mattering.

In no book on introductory probability should problems on tossing coins, rolling dice, and dealing cards be missing. The next examples deal with these sorts of probability problems and use binomial coefficients to solve them.

Example 1.4 A fair coin is tossed 100 times. What is the probability of getting exactly 50 heads?

Solution. Take as sample space all possible sequences of zeros and ones to a length of 100, where a zero stands for tails and a one for heads. The sample space has 2^{100} equiprobable elements. The number of elements having exactly 50 ones is $\binom{100}{50}$. Therefore, the probability of getting exactly 50 heads is

$$\frac{\binom{100}{50}}{2^{100}}.$$

This probability is the ratio of two enormously large numbers and its computation requires special provisions. However, a very accurate approximation to this probability can be given. To this end, consider the general case of $2N$ tosses of a fair coin. Then the probability p_N of getting exactly N heads is

$$p_N = \frac{\binom{2N}{N}}{2^{2N}}.$$

To approximate this probability, we use Stirling's approximation for $n!$. This famous approximation states that

$$n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$$

for sufficiently large n, where the mathematical constant e is the Euler number 2.7182818 . . .. In practice, this approximation is useful for $n \geq 10$. The relative error percentage is about $\frac{100}{12n}\%$. Using Stirling's approximation, we find

$$p_N = \frac{(2N)!}{N!\,N!}\,\frac{1}{2^{2N}} \approx \frac{\sqrt{2\pi \times 2N}\,(2N/e)^{2N}}{\sqrt{2\pi N}\,(N/e)^N\,\sqrt{2\pi N}\,(N/e)^N}\,\frac{1}{2^{2N}}.$$

Canceling out common terms in the denominator and numerator, we get

$$p_N \approx \frac{1}{\sqrt{\pi N}}$$

for N sufficiently large. This approximation is not only very accurate, but also gives insight into how the probability p_N depends on N. The approximate value of p_{50} is 0.07979, while the exact value is 0.07959.

Example 1.5 What is the probability that three different face values each appear twice in a roll of six dice?

Solution. Take as sample space the set of all possible sequences of the face values $1, 2, \ldots, 6$ to a length of 6. The sample space has 6^6 equiprobable elements. There are $\binom{6}{3}$ ways to choose the face values for the three pairs of different face values. There are $\binom{6}{2}$ possible combinations of two dice from the six dice for the first pair, $\binom{4}{2}$ possible combinations of two dice from the remaining four dice for the second pair, and then one combination of two dice remains for the third pair. Thus the probability of getting three pairs of different face values in a roll of six dice is

$$\frac{\binom{6}{3} \times \binom{6}{2} \times \binom{4}{2} \times 1}{6^6} = 0.0386.$$

The next example shows that the choice of the sample space is not always unambiguous.

Example 1.6 A bridge hand in which there is no card higher than a nine is called a Yarborough. What is the probability of a Yarborough when you are randomly dealt 13 cards out of a well-shuffled deck of 52 cards?

Solution. The choice of the sample space depends on whether we care about the order in which the cards are dealt from the deck of 52 cards. If we consider the order in which the 13 cards are dealt as being relevant, then we take a