1 Introduction and review of probability

1.1 Probability models

Stochastic processes constitute a branch of probability theory treating probabilistic systems that evolve in time. There seems to be no very good reason for trying to define stochastic processes precisely, but as we hope will become evident in this chapter, there is a very good reason for trying to be precise about probability itself. Those particular topics in which evolution in time is important will then unfold naturally. Section 1.5 gives a brief introduction to one of the very simplest stochastic processes, the Bernoulli process, and then Chapters 2, 3, and 4 develop three basic stochastic process models which serve as simple examples and starting points for the other processes to be discussed later.

Probability theory is a central field of mathematics, widely applicable to scientific, technological, and human situations involving uncertainty. The most obvious applications are to situations, such as games of chance, in which repeated trials of essentially the same procedure lead to differing outcomes. For example, when we flip a coin, roll a die, pick a card from a shuffled deck, or spin a ball onto a roulette wheel, the procedure is the same from one trial to the next, but the outcome (heads (H) or tails (T) in the case of a coin, 1 to 6 in the case of a die, etc.) varies from one trial to another in a seemingly random fashion.

For the case of flipping a coin, the outcome of the flip could be predicted from the initial position, velocity, and angular momentum of the coin and from the nature of the surface on which it lands. Thus, in one sense, a coin flip is deterministic rather than random and the same can be said for the other examples above. When these initial conditions are unspecified, however, as when playing these games, the outcome can again be viewed as random in some intuitive sense.

Many scientific experiments are similar to games of chance in the sense that multiple trials of apparently the *same* procedure lead to results that *vary* from one trial to another. In some cases, this variation is due to slight variations in the experimental procedure, in some it is due to noise, and in some, such as in quantum mechanics, the randomness is generally believed to be fundamental. Similar situations occur in many types of systems, especially those in which noise and random delays are important. Some of these systems, rather than being repetitions of a common basic procedure, are systems that evolve over time while still containing a sequence of underlying similar random occurrences.

This intuitive notion of randomness, as described above, is a very special kind of uncertainty. Rather than involving a lack of understanding, it involves a type of uncertainty that can lead to probabilistic models with precise results. As in any scientific field, the models might or might not correspond to reality very well, but when they do correspond to reality, there is the sense that the situation is completely understood, while still being random.

For example, we all feel that we understand flipping a coin or rolling a die, but still accept randomness in each outcome. The theory of probability was initially developed particularly to give precise and quantitative meaning to these types of situations. The remainder of this section introduces this relationship between the precise view of probability theory and the intuitive view as used in applications and everyday language.

After this introduction, the following sections of this chapter review probability theory as a mathematical discipline, with a special emphasis on the laws of large numbers. In the final section, we use the theory and the laws of large numbers to obtain a fuller understanding of the relationship between theory and the real world.[1]

Probability theory, as a mathematical discipline, started to evolve in the seventeenth century and was initially focused on games of chance. The importance of the theory grew rapidly, particularly in the twentieth century, and it now plays a central role in risk assessment, statistics, data networks, operations research, information theory, control theory, theoretical computer science, quantum theory, game theory, neurophysiology, and many other fields.

The core concept in probability theory is that of a *probability model*. Given the extent of the theory, both in mathematics and in applications, the simplicity of probability models is surprising. The first component of a probability model is a *sample space*, which is a *set* whose elements are called *sample points* or *outcomes*. Probability models are particularly simple in the special case where the sample space is finite, and we consider only this case in the remainder of this section. The second component of a probability model is a class of *events*, which can be considered for now simply as the class of all subsets of the sample space. The third component is a *probability measure*, which can be regarded for now as the assignment of a non-negative number to each outcome, with the restriction that these numbers must sum to 1 over the sample space. The probability of an event is the sum of the probabilities of the outcomes comprising that event.

These probability models play a dual role. In the first, the many known results about various classes of models, and the many known relationships between models, constitute the essence of probability theory. Thus one often studies a model not because of any relationship to the real world, but simply because the model provides a building block or example useful for the theory and thus ultimately useful for other models. In the other role, when probability theory is applied to some game, experiment, or other situation

[1] It would be appealing to show how probability theory evolved from real-world random situations, but probability theory, like most mathematical theories, has evolved from complex interactions between theoretical developments and initially oversimplified models of real situations. The successes and flaws of such models lead to refinements of the models and the theory, which in turn suggest applications to totally different fields.

involving randomness, a probability model is used to represent the experiment (in what follows, we refer to all of these random situations as experiments).

For example, the standard probability model for rolling a die uses $\{1, 2, 3, 4, 5, 6\}$ as the sample space, with each possible outcome having probability 1/6. An *odd* result, i.e., the subset $\{1, 3, 5\}$, is an example of an event in this sample space, and this event has probability 1/2. The correspondence between model and actual experiment seems straightforward here. Both have the same set of outcomes and, given the symmetry between faces of the die, the choice of equal probabilities seems natural. Closer inspection, however, reveals an important difference between the model and the actual rolling of a die.

The model above corresponds to a single roll of a die, with a probability defined for each possible outcome. In a real-world experiment where a single die is rolled, one of the six faces, say face k comes up, but there is no *observable* probability for k.

Our intuitive notion of rolling dice, however, involves an experiment with n consecutive rolls of a die. There are then 6^n possible outcomes, one for each possible n-tuple of individual die outcomes. As reviewed in subsequent sections, the standard probability model for this repeated-roll experiment is to assign probability 6^{-n} to each possible n-tuple, which leads to a probability $\binom{n}{m}(1/6)^m(5/6)^{n-m}$ that the face k comes up on m of the n rolls, i.e., that the relative frequency of face k is m/n. The distribution of these relative frequencies is increasingly clustered around $1/6$ as n is increased. Thus if a real-world experiment for tossing n dice is reasonably modeled by this probability model, we would also expect the relative frequency to be close to 1/6 for large n. This relationship through relative frequencies in a repeated experiment helps overcome the non-observable nature of probabilities in the real world.

1.1.1 The sample space of a probability model

An *outcome* or *sample point* in a probability model corresponds to a complete result (with all detail specified) of the experiment being modeled. For example, a game of cards is often appropriately modeled by the arrangement of cards within a shuffled 52-card deck, thus giving rise to a set of 52! outcomes (incredibly detailed, but trivially simple in structure), even though the entire deck might not be played in one trial of the game. A poker hand with four aces is an *event* rather than an *outcome* in this model, since many arrangements of the cards can give rise to four aces in a given hand. The possible outcomes in a probability model (and in the experiment being modeled) are mutually exclusive and collectively constitute the entire sample space (space of possible outcomes). An outcome ω is often called a *finest grain* result of the model in the sense that a singleton event $\{\omega\}$ containing only ω clearly contains no proper subsets. Thus events (other than singleton events) typically give only partial information about the result of the experiment, whereas an outcome fully specifies the result.

In choosing the sample space for a probability model of an experiment, we often omit details that appear irrelevant for the purpose at hand. Thus in modeling the set of outcomes for a coin toss as $\{H, T\}$, we ignore the type of coin, the initial velocity and angular momentum of the toss, etc. We also omit the rare possibility that the coin comes

to rest on its edge. Sometimes, conversely, the sample space is enlarged beyond what is relevant in the interest of structural simplicity. An example is the above use of a shuffled deck of 52 cards.

The choice of the sample space in a probability model is similar to the choice of a mathematical model in any branch of science. That is, one simplifies the physical situation by eliminating detail of little apparent relevance. One often does this in an iterative way, using a very simple model to acquire initial understanding, and then successively choosing more detailed models based on the understanding from earlier models.

The mathematical theory of probability views the sample space simply as an abstract set of elements, and from a strictly mathematical point of view, the idea of doing an experiment and getting an outcome is a distraction. For visualizing the correspondence between the theory and applications, however, it is better to view the abstract set of elements as the set of possible outcomes of an idealized experiment in which, when the idealized experiment is performed, one and only one of those outcomes occurs. The two views are mathematically identical, but it will be helpful to refer to the first view as a probability model and the second as an idealized experiment. In applied probability texts and technical articles, these idealized experiments, rather than real-world situations, are often the primary topic of discussion.[2]

1.1.2 Assigning probabilities for finite sample spaces

The word *probability* is widely used in everyday language, and most of us attach various intuitive meanings[3] to the word. For example, everyone would agree that something virtually impossible should be assigned a probability close to 0 and something virtually certain should be assigned a probability close to 1. For these special cases, this provides a good rationale for choosing probabilities. The meaning of *virtually* and *close to* are slightly unclear at the moment, but if there is some implied limiting process, we would all agree that, in the limit, certainty and impossibility correspond to probabilities 1 and 0 respectively.

Between virtual impossibility and certainty, if one outcome appears to be closer to certainty than another, its probability should be correspondingly greater. This intuitive notion is imprecise and highly subjective; it provides little rationale for choosing numerical probabilities for different outcomes, and, even worse, little rationale justifying that probability models bear any precise relation to real-world situations.

Symmetry can often provide a better rationale for choosing probabilities. For example, the symmetry between H and T for a coin, or the symmetry between the six faces of a die, motivates assigning equal probabilities, 1/2 each for H and T and 1/6 each for the six faces of a die. This is reasonable and extremely useful, but there is no completely convincing reason for choosing probabilities based on symmetry.

[2] This is not intended as criticism, since we will see that there are good reasons to concentrate initially on such idealized experiments. However, readers should always be aware that modeling errors are the major cause of misleading results in applications of probability, and thus modeling must be seriously considered before using the results.

[3] It is popular to try to define probability by likelihood, but this is unhelpful since the words are essentially synonyms.

Another approach is to perform the experiment many times and choose the probability of each outcome as the relative frequency of that outcome (i.e., the number of occurrences of that outcome divided by the total number of trials). Experience shows that the relative frequency of an outcome often approaches a limiting value with an increasing number of trials. Associating the probability of an outcome with that limiting relative frequency is certainly close to our intuition and also appears to provide a testable criterion between model and real world. This criterion is discussed in Sections 1.8.1 and 1.8.2 and provides a very concrete way to use probabilities, since it suggests that the randomness in a single trial tends to disappear in the aggregate of many trials. Other approaches to choosing probability models will be discussed later.

1.2 The axioms of probability theory

As the applications of probability theory became increasingly varied and complex during the twentieth century, the need arose to put the theory on a firm mathematical footing. This was accomplished by an axiomatization of the theory, successfully carried out by the great Russian mathematician A. N. Kolmogorov [18] in 1932. Before stating and explaining these axioms of probability theory, the following two examples explain why the simple approach of the last section, assigning a probability to each sample point, often fails with infinite sample spaces.

Example 1.2.1 Suppose we want to model the phase of a sine wave, where the phase is viewed as being 'uniformly distributed' between 0 and 2π. If this phase is the only quantity of interest, it is reasonable to choose a sample space consisting of the set of real numbers between 0 and 2π. There are uncountably[4] many possible phases between 0 and 2π, and with any reasonable interpretation of uniform distribution, one must conclude that each sample point has probability 0. Thus, the simple approach of the last section leads us to conclude that any event in this space with a finite or countably infinite set of sample points should have probability 0. That simple approach does not help in finding the probability, say, of the interval $(0, \pi)$.

For this example, the appropriate view is the one taken in all elementary probability texts, namely to assign a *probability density* $1/(2\pi)$ to the phase. The probability of an event can then usually be found by integrating the density over that event. Useful as densities are, however, they do not lead to a general approach over arbitrary sample spaces.[5]

[4] A set is uncountably infinite if it is infinite and its members cannot be put into one-to-one correspondence with the positive integers. For example, the set of real numbers over some interval such as $(0, 2\pi)$ is uncountably infinite. The Wikipedia article on countable sets provides a friendly introduction to the concepts of countability and uncountability.

[5] It is possible to avoid the consideration of infinite sample spaces here by quantizing the possible phases. This is analogous to avoiding calculus by working only with discrete functions. Both usually result in both artificiality and added complexity.

Example 1.2.2 Consider an infinite sequence of coin tosses. The usual probability model is to assign probability 2^{-n} to each possible initial n-tuple of individual outcomes. Then in the limit $n \to \infty$, the probability of any given sequence is 0. Again, expressing the probability of an event involving infinitely many tosses as a sum of individual sample-point probabilities does not work. The obvious approach (which we often adopt for this and similar situations) is to evaluate the probability of any given event as an appropriate limit, as $n \to \infty$, of the outcome from the first n tosses.

We will later find a number of situations, even for this almost trivial example, where working with a finite number of elementary experiments and then going to the limit is very awkward. One example, to be discussed in detail later, is the strong law of large numbers (SLLN). This law looks directly at events consisting of infinite length sequences and is best considered in the context of the axioms to follow.

Although appropriate probability models can be generated for simple examples such as those above, there is a need for a consistent and general approach. In such an approach, rather than assigning probabilities to sample points, which are then used to assign probabilities to events, *probabilities must be associated directly with events.* The axioms to follow establish consistency requirements between the probabilities of different events. The axioms, and the corollaries derived from them, are consistent with one's intuition, and, for finite sample spaces, are consistent with our earlier approach. Dealing with the countable unions of events in the axioms will be unfamiliar to some students, but will soon become both familiar and consistent with intuition.

The strange part of the axioms comes from the fact that defining the class of events as the collection of *all* subsets of the sample space is usually inappropriate when the sample space is uncountably infinite. What is needed is a class of events that is large enough that we can almost forget that some very strange subsets are excluded. This is accomplished by having two simple sets of axioms, one defining the class of events,[6] and the other defining the relations between the probabilities assigned to these events. In this theory, all events have probabilities, but those truly weird subsets that are not events do not have probabilities. This will be discussed more after giving the axioms for events.

The axioms for events use the standard notation of set theory. Let Ω be the sample space, i.e., the set of all sample points for a given experiment. It is assumed throughout that Ω is non-empty. The events are subsets of the sample space. The union of n subsets (events) $A_1, A_2, \ldots, A_n$ is denoted by either $\bigcup_{i=1}^{n} A_i$ or $A_1 \bigcup \cdots \bigcup A_n$, and consists of all points in at least one of $A_1, A_2, \ldots, A_n$. Similarly, the intersection of these subsets is denoted by either $\bigcap_{i=1}^{n} A_i$ or[7] $A_1 A_2 \cdots A_n$ and consists of all points in all of $A_1, A_2, \ldots, A_n$.

A *sequence* of events is a collection of events in one-to-one correspondence with the positive integers, i.e., $A_1, A_2, \ldots$ ad infinitum. A countable union, $\bigcup_{i=1}^{\infty} A_i$ is the set of

[6] A class of elements satisfying these axioms is called a σ-algebra or, less commonly, a σ-field.

[7] Intersection is also sometimes denoted as $A_1 \bigcap \cdots \bigcap A_n$, but is usually abbreviated as $A_1 A_2 \cdots A_n$.

points in one or more of $A_1, A_2, \ldots$. Similarly, a countable intersection $\bigcap_{i=1}^{\infty} A_i$ is the set of points in all of $A_1, A_2, \ldots$. Finally, the complement A^c of a subset (event) A is the set of points in Ω but not A.

1.2.1 Axioms for events

Given a sample space Ω, the class of subsets of Ω that constitute the set of events satisfies the following axioms:

1. Ω is an event.
2. For every sequence of events $A_1, A_2, \ldots$, the union $\bigcup_{n=1}^{\infty} A_n$ is an event.
3. For every event A, the complement A^c is an event.

There are a number of important corollaries of these axioms. First, the empty set $\emptyset$ is an event. This follows from Axioms 1 and 3, since $\emptyset = \Omega^c$. The empty set does not correspond to our intuition about events, but the theory would be extremely awkward if it were omitted. Second, every finite union of events is an event. This follows by expressing $A_1 \bigcup \cdots \bigcup A_n$ as $\bigcup_{i=1}^{\infty} A_i$, where $A_i = \emptyset$ for all $i > n$. Third, every finite or countable intersection of events is an event. This follows from De Morgan's law,

$$\left[\bigcup_n A_n\right]^c = \bigcap_n A_n^c.$$

Although we will not make a big fuss about these axioms in the rest of the text, we will be careful to use only complements and countable unions and intersections in our analysis. Thus subsets that are not events will not arise.

Note that the axioms do not say that all subsets of Ω are events. In fact, there are many rather silly ways to define classes of events that obey the axioms. For example, the axioms are satisfied by choosing only the universal set Ω and the empty set $\emptyset$ to be events. We shall avoid such trivialities by assuming that for each sample point ω, the singleton subset $\{\omega\}$ is an event. For finite sample spaces, this assumption, plus the axioms above, imply that all subsets are events.

For uncountably infinite sample spaces, such as the sinusoidal phase above, this assumption, plus the axioms above, still leaves considerable freedom in choosing a class of events. As an example, the class of all subsets of Ω satisfies the axioms but surprisingly does not allow the probability axioms to be satisfied in any sensible way. How to choose an appropriate class of events requires an understanding of measure theory which would take us too far afield for our purposes. Thus we neither assume nor develop measure theory here.[8]

From a pragmatic standpoint, we start with the class of events of interest, such as those required to define the random variables (rv s) needed in the problem. That class is then extended so as to be closed under complementation and countable unions. Measure theory shows that this extension is possible.

[8] There is no doubt that measure theory is useful in probability theory, and serious students of probability should certainly learn measure theory at some point. For application-oriented people, however, it seems advisable to acquire more insight and understanding of probability, at a graduate level, before concentrating on the abstractions and subtleties of measure theory.

1.2.2 Axioms of probability

Given any sample space Ω and any class of events $\mathcal{E}$ satisfying the axioms of events, a probability rule is a function $\Pr\{\cdot\}$ mapping each $A \in \mathcal{E}$ to a (finite[9]) real number in such a way that the following three probability axioms[10] hold:

1. $\Pr\{\Omega\} = 1$.
2. For every event A, $\Pr\{A\} \geq 0$.
3. The probability of the union of any sequence $A_1, A_2, \ldots$ of disjoint[11] events is given by

$$\Pr\left\{\bigcup_{n=1}^{\infty} A_n\right\} = \sum_{n=1}^{\infty} \Pr\{A_n\}, \tag{1.1}$$

where $\sum_{n=1}^{\infty} \Pr\{A_n\}$ is shorthand for $\lim_{m\to\infty} \sum_{n=1}^{m} \Pr\{A_n\}$.

The axioms imply the following useful corollaries:

$$\Pr\{\emptyset\} = 0. \tag{1.2}$$

$$\Pr\left\{\bigcup_{n=1}^{m} A_n\right\} = \sum_{n=1}^{m} \Pr\{A_n\} \qquad \text{for } A_1, \ldots, A_m \text{ disjoint.} \tag{1.3}$$

$$\Pr\{A^c\} = 1 - \Pr\{A\} \qquad \text{for all } A. \tag{1.4}$$

$$\Pr\{A\} \leq \Pr\{B\} \qquad \text{for all } A \subseteq B. \tag{1.5}$$

$$\Pr\{A\} \leq 1 \qquad \text{for all } A. \tag{1.6}$$

$$\sum_n \Pr\{A_n\} \leq 1 \qquad \text{for } A_1, A_2, \ldots \text{ disjoint.} \tag{1.7}$$

$$\Pr\left\{\bigcup_{n=1}^{\infty} A_n\right\} = \lim_{m\to\infty} \Pr\left\{\bigcup_{n=1}^{m} A_n\right\}. \tag{1.8}$$

$$\Pr\left\{\bigcup_{n=1}^{\infty} A_n\right\} = \lim_{n\to\infty} \Pr\{A_n\} \qquad \text{for } A_1 \subseteq A_2 \subseteq \cdots. \tag{1.9}$$

$$\Pr\left\{\bigcap_{n=1}^{\infty} A_n\right\} = \lim_{n\to\infty} \Pr\{A_n\} \qquad \text{for } A_1 \supseteq A_2 \supseteq \cdots. \tag{1.10}$$

To verify (1.2), consider a sequence of events, $A_1, A_2, \ldots$ for which $A_n = \emptyset$ for each n. These events are disjoint since $\emptyset$ contains no outcomes, and thus has no outcomes in common with itself or any other event. Also, $\bigcup_n A_n = \emptyset$ since this union contains no outcomes. Axiom 3 then says that

$$\Pr\{\emptyset\} = \lim_{m\to\infty} \sum_{n=1}^{m} \Pr\{A_n\} = \lim_{m\to\infty} m\Pr\{\emptyset\}.$$

Since $\Pr\{\emptyset\}$ is a real number, this implies that $\Pr\{\emptyset\} = 0$.

To verify (1.3), apply Axiom 3 to the disjoint sequence $A_1, \ldots, A_m, \emptyset, \emptyset, \ldots$.

To verify (1.4), note that $\Omega = A \bigcup A^c$. Then apply (1.3) to the disjoint sets A and A^c.

[9] The word *finite* is redundant here, since the set of real numbers, by definition, does not include $\pm\infty$. The set of real numbers with $\pm\infty$ appended, is called the *extended* set of real numbers.

[10] Sometimes finite additivity, (1.3), is included as an additional axiom. This inclusion is quite intuitive and avoids the technical and somewhat peculiar proofs given for (1.2) and (1.3).

[11] Two sets or events A_1, A_2 are disjoint if they contain no common events, i.e., if $A_1 A_2 = \emptyset$. A collection of sets or events are disjoint if all pairs are disjoint.

To verify (1.5), note that if $A \subseteq B$, then $B = A \bigcup (B-A)$, where $B-A$ is an alternative way to write $B \bigcap A^c$. We see then that A and $B-A$ are disjoint, so from (1.3),

$$\Pr\{B\} = \Pr\left\{A \bigcup (B-A)\right\} = \Pr\{A\} + \Pr\{B-A\} \geq \Pr\{A\},$$

where we have used Axiom 2 in the last step.

To verify (1.6) and (1.7), first substitute Ω for B in (1.5) and then substitute $\bigcup_n A_n$ for A.

Finally, (1.8) is established in Exercise 1.2(e), and (1.9) and (1.10) are simple consequences of (1.8).

The axioms specify the probability of any *disjoint* union of events in terms of the individual event probabilities, but what about a finite or countable union of arbitrary events? Exercise 1.2(c) shows that in this case, (1.3) can be generalized to

$$\Pr\left\{\bigcup_{n=1}^{m} A_n\right\} = \sum_{n=1}^{m} \Pr\{B_n\}, \tag{1.11}$$

where $B_1 = A_1$ and for each $n > 1$, $B_n = A_n - \bigcup_{m=1}^{n-1} A_m$ is the set of points in A_n but not in any of the sets $A_1, \ldots, A_{n-1}$. That is, the sets B_n are disjoint. The probability of a countable union of disjoint sets is then given by (1.8). In order to use this, one must know not only the event probabilities for $A_1, A_2 \ldots$, but also the probabilities of their intersections. The union bound, which is derived in Exercise 1.2(d), depends only on the individual event probabilities, and gives the following frequently useful upper bound on the union probability.

$$\Pr\left\{\bigcup_n A_n\right\} \leq \sum_n \Pr\{A_n\} \qquad \text{(union bound)}. \tag{1.12}$$

1.3 Probability review

1.3.1 Conditional probabilities and statistical independence

Definition 1.3.1 *For any two events A and B in a probability model, the **conditional probability** of A, conditional on B, is defined if $\Pr\{B\} > 0$ by*

$$\Pr\{A|B\} = \Pr\{AB\} / \Pr\{B\}. \tag{1.13}$$

To motivate this definition, consider a discrete experiment in which we make a partial observation B (such as the result of a given medical test on a patient) but do not observe the complete outcome (such as whether the patient is sick and the outcome of other tests). The event B consists of all the sample points with the given outcome of the given test. Now let A be an arbitrary event (such as the event that the patient is sick). The conditional probability, $\Pr\{A|B\}$ is intended to represent the probability of A from the observer's viewpoint.

For the observer, the sample space can now be viewed as the set of sample points in B, since only those sample points are now possible. For any event A, only the event AB, i.e., the original set of sample points in A that are also in B, is relevant, but the probability of

A in this new sample space should be scaled up from $\Pr\{AB\}$ to $\Pr\{AB\}/\Pr\{B\}$, i.e., to $\Pr\{A|B\}$.

With this scaling, the set of events conditional on B becomes a probability space, and it is easily verified that all the axioms of probability theory are satisfied for this conditional probability space. Thus all known results about probability can also be applied to such conditional probability spaces.

Another important aspect of the definition in (1.13) is that it maintains consistency between the original probability space and this new conditional space in the sense that for any disjoint events, $A_1, A_2, \ldots$, and any event B with $\Pr\{B\} > 0$,

$$\Pr\left\{\left(\bigcup_n A_n\right) \mid B\right\} = \sum_n \Pr\{A_n \mid B\}\,.$$

This means that we can easily move back and forth between unconditional and conditional probability spaces.

The intuitive statements about partial observations and probabilities from the standpoint of an observer are helpful in reasoning probabilistically, but sometimes cause confusion. For example, Bayes' law, in the form

$$\Pr\{A|B\}\Pr\{B\} = \Pr\{B|A\}\Pr\{A\}\,,$$

is an immediate consequence of the definition of conditional probability in (1.13). However, if we can only interpret $\Pr\{A|B\}$ when B is 'observed' or occurs 'before' A, then we cannot interpret $\Pr\{B|A\}$ and $\Pr\{A|B\}$ together. This caused immense confusion in probabilistic arguments before the axiomatic theory and clean definitions based on axioms were developed.

Definition 1.3.2 *Two events, A and B, are **statistically independent** (or, more briefly, independent) if*

$$\Pr\{AB\} = \Pr\{A\}\Pr\{B\}\,.$$

For $\Pr\{B\} > 0$, this is equivalent to $\Pr\{A|B\} = \Pr\{A\}$. This latter form often corresponds to a more intuitive view of independence, since it says that A and B are independent if the observation of B does not change the observer's probability of A.

The notion of independence is of vital importance in defining, and reasoning about, probability models. We will see many examples where very complex systems become very simple, both in terms of intuition and analysis, when appropriate quantities are modeled as statistically independent. An example will be given in the next subsection where repeated independent experiments are used to understand arguments about relative frequencies.

Often, when the assumption of independence is unreasonable, it is reasonable to assume conditional independence, where A and B are said to be *conditionally independent* given C if $\Pr\{AB|C\} = \Pr\{A|C\}\Pr\{B|C\}$. Most of the stochastic processes to be studied here are characterized by various forms of independence or conditional independence.

For more than two events, the definition of statistical independence is a little more complicated.