PART I

INTRODUCTION

1
The meaning of 'probability'

Probability theory has a long, eventful, and still not fully settled history. As pointed out in [63]: 'For all human history, people have invented methods for coming to terms with the seemingly unpredictable vicissitudes of existence ... Oracles, amulets, and incantations belonged to the indispensable techniques for interpreting and influencing the fate of communities and individuals alike ... In the place of superstition there was to be calculation – a project aiming at nothing less than the rationalization of fortune. From that moment on, there was no more talk of fortune but instead of this atrophied cousin: chance.'

The only consistent mathematical way to handle chance, or rather probability, is provided by the rules of (Bayesian) probability theory. But what does the notion 'probability' really mean? Although it might appear, at first sight, as obvious, it actually has different connotations and definitions, which will be discussed in the following sections.

For the sake of a smooth introduction to probability theory, we will forego a closer definition of some technical terms, as long as their colloquial meaning suffices for understanding the concepts. A precise definition of these terms will be given in a later section.

1.1 Classical definition of 'probability'

The first quantitative definition of the term 'probability' appears in the work of Blaise Pascal (1623–1662) and Pierre de Fermat (1601–1665). Antoine Gombauld Chevalier de Méré, Sieur de Baussay (1607–1685) pointed out to them that '...mathematics does not apply to real life'. For this nobleman 'real life' at that time meant gambling. He was especially interested in the odds of having at least once the value '6' in four rolls of a die, which was of importance in a common game of chance at the time. The analysis of this problem will be discussed along with equation (4.4) [p. 49]. Pascal and Fermat studied this and related problems and developed the basic concepts underlying classical probability theory that are still used today.

Definition 1.1 (Classical definition of probability) *The probability for the occurrence of a random event is defined as the ratio of the number g of favourable outcomes for the event to the total the number m of possibilities.*

Classical definition of 'probability'

$$P = \frac{g}{m}. \tag{1.1}$$

Example 1.1 *The probability for selecting a card of the suit 'spades' from a deck of bridge cards is* $P = \frac{13}{52} = \frac{1}{4}$.

Based on the classical definition we can immediately derive the basic rules of probability theory. Let A and B be arbitrary events and $\vee$ stand for the logical 'or' and $\wedge$ for the logical 'and' (please see Table C.1 for a list of symbols used in the book). Then

$$P(A \vee B) = \frac{n_A + n_B - n_{A\wedge B}}{N} \qquad = P(A) + P(B) - P(A \wedge B); \tag{1.2a}$$

$$P(N) = \frac{0}{N} = 0 \qquad N\text{: impossible event;} \tag{1.2b}$$

$$P(E) = \frac{N}{N} = 1 \qquad E\text{: certain event;} \tag{1.2c}$$

$$0 \leq P(A) \leq 1 \qquad \text{follows from the definition.} \tag{1.2d}$$

We are also prompted to introduce the conditional probability

$$P(A|B) := \frac{n_{A\wedge B}}{n_B} = \frac{P(A \wedge B)}{P(B)}, \tag{1.2e}$$

which is the probability for event A provided B is true. Within the classical definition this can be considered as a kind of pre-selection. Of all possible events only those n_B events are considered which are compatible with the condition implied by B. Of these n_B events only those $n_{A\wedge B}$ are considered favourable which in addition are compatible with the condition implied by A.

Definition 1.2 (Exclusive events) *Events are said to be exclusive if the occurrence of any one of them implies the non-occurrence of any of the remaining events, i.e.* $A \wedge B = N$.

Definition 1.3 (Complementary events) *An event* $\overline{A}$ *is said to be complementary to* A *if*

$$\overline{A} \vee A = E \qquad \textit{and} \qquad \overline{A} \wedge A = N.$$

The general sum rule equation (1.2a) simplifies for exclusive events to

$$P(A \vee B) = P(A) + P(B) \tag{1.3}$$

and thus the relation for complementary events follows:

$$P(\overline{A}) = 1 - P(A). \tag{1.4}$$

These ideas can be generalized to continuous problems. Consider, for example, a square whose edge length is L. Entirely within the square shall be a circle of radius r. Now we generate points inside the square at random; no area element is distinguished from the others. We divide the square into a fine grid of squares and the random points are classified according to which grid point they land in. Now we can apply equation (1.1) again. The total number of possible outcomes is the number of grid points N. The favourable number of outcomes is equal to the number of grid points whose centre lies within the circle. In the limit $N \to \infty$ the sought-for probability is given by the ratio of the corresponding areas. Or in general, we have

$$P = \frac{\text{volume corresponding to favourable events}}{\text{total volume}}. \tag{1.5}$$

For the problem under consideration the probability that a random point lies inside the circle is

$$P = \frac{\pi r^2}{L^2}.$$

Alternatively, we can solve the inverse problem and infer r, if L is given and if we know that of N random points n are inside the circle. As a matter of fact, this is an elementary example of Monte Carlo integration. Both topics, inverse reasoning and Monte Carlo integration, will be discussed in great detail in later chapters.

The classical definition of 'probability' was developed further by Jacob Bernoulli (1654–1705) in his book *Ars Conjectandi* (published posthumously in 1713). This book contains seminal contributions to the theory of probability, among others an extensive discussion on the 'true' meaning of the term probability: *Probability is a measure of certainty.* Bernoulli already distinguished between prior and posterior probabilities. Later, Pierre-Simon Laplace (1749–1827) systematized and extended the field of probability theory. He already applied it to inverse reasoning (e.g. given the street is wet, what is the probability that it has rained). The formula his reasoning was based upon had been derived by Reverend Thomas Bayes (1702–1761). This formula, nowadays known as Bayes' theorem, was published posthumously by Richard Price [10] in 1764. It was quite normal that clergymen at the time were also (amateur) scientists [26]. Remarkable, though, is the result itself. It represents, as we will see, the only consistent solution for inverse problems and it was revealed at a time when inverse conclusions had been drawn based on bizarre logic intertwined with superstition. The original presentation of Bayes' ideas [10], however, was not very revealing and Laplace was the first to restate the theorem in the form known today. In hindsight, it is a very simple application of the product rule.

Laplace applied probability theory to problems in celestial mechanics, games of chance, the needle problem of Buffon, court cases, and many more [123]. He also introduced the principle of indifference to assign prior probabilities.

It was already known to Bernoulli that the definition given in equation (1.1) is not unique, as the total number m of all events and the number g of favourable events are sometimes

ambiguous. Consider the following 'two-dice paradox': Two dice are rolled and we are interested in the probability for the sum of the two face values being seven. There are several conceivable approaches:

(a) We consider as possible outcomes the 11 different sums of the face values $(2, 3, \ldots, 12)$. Of these 11 only one result (7) is favourable. Resulting in $P = 1/11$.
(b) The possible outcomes are the ordered pairs of the face values, i.e. $(1, 1), (1, 2), \ldots, (1, 6), (2, 2), (2, 3), \ldots, (6, 6)$. Here we do not distinguish between the two dice. Hence there are $6 + 5 + 4 + 3 + 2 + 1 = 21$ possible pairs out of which three are favourable $((1, 6), (2, 5), (3, 4))$. Resulting in $P = 1/7$.
(c) Now we also take into account which die displayed which face value. Then we end up with $N = 36$ and $g = 6$, the favourable events being $(1, 6), (6, 1), (2, 5), (5, 2), (3, 4), (4, 3)$. Resulting in $P = 1/6$.

In this example it appears obvious that the last approach is the correct one. However, as we will see later, there are also situations (especially in quantum physics) where the second approach is to apply. This immediately implies that a refined definition of probability is necessary.

Definition 1.4 (Refined classical definition of 'probability') *The probability for an event is given by the ratio of the number of favourable events to the total number of events, if all events are equally likely.*

This definition is based on circular reasoning and assumes that it is possible to assign 'prior probabilities' to elementary events, those that cannot be decomposed further. Nevertheless, it separates the rules for manipulating probabilities from the assignment of values to the 'prior probabilities' of elementary events. In many cases, it is indeed clear how to assign probabilities to the elementary events and to apply successfully the classical definition of probability. Typical examples are games of chance. The classical definition even forms the basis for statistical physics, as it is presented in most textbooks.

In order to assign probabilities, two principles were suggested early on, the 'Principle of Insufficient Reasoning' by Jacob Bernoulli and the 'Principle of Indifference' by Pierre-Simon Laplace. Both principles state that if there are n possibilities which are in principle indistinguishable, then each possibility should be assigned an equal probability, because a reordering would not alter the situation. Although these principles apply to many situations, there are counterexamples known in statistical physics (e.g. the example given above) and the principles do not generalize easily to continuous problems. A famous example is the Bertrand paradox (1888), which was resolved only in 1968 by E. T. Jaynes (1922–1998).

1.1.1 Bertrand paradox

Suppose straight lines are randomly drawn across a circle. What is the probability that the distance of a line from the centre of the circle is smaller than half of the radius r? Without loss of generality, we set $r = 1$ in suitable length units. To begin with, we have

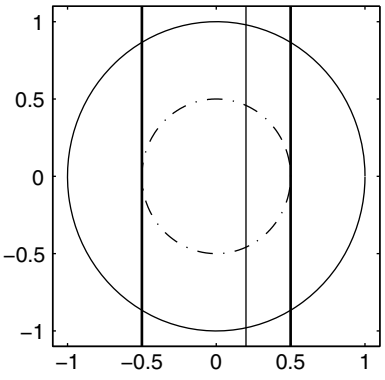

Figure 1.1 Illustration of approach (1) for the Bertrand paradox.

to generalize the classical definition in equation (1.1) to continuous variables. Let the real variable x describe the random event and by construction, all allowed events correspond to $x \in I$ with the size of I being $|I| = L$. The favourable events are given by $x \in I_f$, with $|I_f| = L_f$. Then, according to equation (1.5), the probability that x is in I_f is

$$P = \frac{L_f}{L}. \tag{1.6}$$

Several representations are conceivable. In the following we discuss three possibilities, all apparently valid.

(1) The distance x to the centre of the circle can be assumed to take a value between 0 and 1, i.e. $I = [0, 1]$ and $L = 1$. The interval of favourable events is $I_f = [0, \frac{1}{2}]$ and has $L = \frac{1}{2}$. So, we end up with a probability $P = 1/2$. See Figure 1.1.
(2) A different approach is illustrated in Figure 1.2: Regarding the angle between the straight line and the tangent to the circle as uniformly distributed, the favourable angles are in the range $I_f = [0, \frac{\pi}{3}]$ compared with $I = [0, \pi]$. The resulting probability is therefore $P = 1/3$.
(3) Another possibility is to consider the area A of the concentric disc touching the straight line (see Figure 1.3). If we consider this area to be equally distributed within $I = [0, \pi]$, and taking into account that the favourable area is limited to $I = [0, \frac{\pi}{4}]$, then the probability of interest is $P = 1/4$.

How can a seemingly well-posed problem yield different answers? At the heart of the Bertrand paradox is the problem of describing ignorance about continuous degrees of freedom. Suppose we want to make a statement about a quantity x which may take real values in $I = [0, 1]$. It appears reasonable to use the classical definition of equation (1.6) and assign the probability that x is within $I_f = [x, x + dx]$ as

$$P(x \in (x, x + dx)) = \frac{L_f}{L} = \frac{dx}{1} = p_x(x)dx.$$

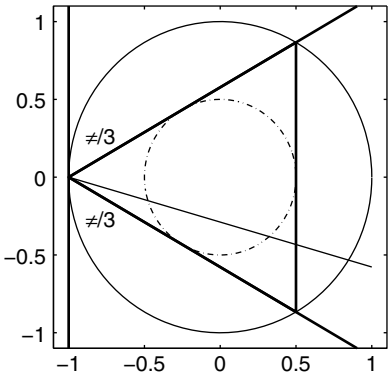

Figure 1.2 Illustration of approach (2) for the Bertrand paradox.

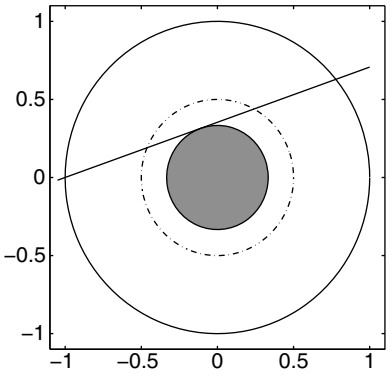

Figure 1.3 Illustration of approach (3) for the Bertrand paradox.

Here, $p_x(x) = 1$ is a 'probability density' which will be treated in more detail in Section 11.1 [p. 178]. Ignorance, or lack of knowledge, results in a uniform assignment. However, now the probability density of x^n is no longer uniform but (as we will see in Chapter 10 [p. 165]) is given by

$$p_z(z = x^n) = p_x(x) \left| \frac{dx}{dz} \right| = \frac{1}{n} z^{\frac{1}{n}-1}.$$

Therefore, the probability density for the variable z has a maximum at $z = 0$. Thus, a uniform probability density for x corresponds to a non-uniform density for the transformed variable z. Quite generally speaking, nonlinear transformations of continuous quantities may cause problems. In which representation are events equally probable (e.g. length, area, volume) and how is complete ignorance best represented? We recognize a similarity to the 'two-dice paradox' but here the situation is much more complex. The problem of 'prior probability assignment' is tackled in Part II [p. 164]. The theory of

'transformation invariance' provides a principled solution to this problem. In the same context, the important concept of 'maximum entropy prior' probabilities is introduced.

Despite all these problems, the classical definition of probability was used for several centuries and is still adequate for many problems, especially in combinatorics. In such problems, there is a close relation between probabilities and relative frequencies of the occurrence of events. Bernoulli was one of the first to analyse this not immediately obvious relation. Let us consider the following example: The probability for rolling an even-face number using a standard die is 1/2. Then, how many even numbers will be obtained in N rolls? We will discuss later in this chapter and in Section 1.3 Bernoulli's 'law of large numbers', which says that the relative frequency approaches the intrinsic probability for $N \to \infty$. However, much more relevant for most applications is the inverse problem: Given a sample of finite size N and n occurrences of the event of interest, what can be inferred about the underlying probability?

1.2 Statistical definition of 'probability'

In order to avoid the need to specify 'prior probabilities,' R. L. Ellis (1772–1842), G. Boole (1815–1864), J. Venn (1834–1923) and R. von Mises (1883–1953) pursued a different direction. Based on Bernoulli's 'law of large numbers', they introduced the following statistical definition.

Definition 1.5 (Statistical definition of probability) *An event A happens at random. The probability for the event is defined by the relative frequency*

$$P(A) = \lim_{N\to\infty} \frac{n}{N} \tag{1.7}$$

that the event occurs n times in N trials, for the limit $N \to \infty$.

In the statistical definition, 'probability' is considered as an intrinsic property of the object under investigation, which is only accessible by an experiment (samples) of infinite size. Since this cannot be realized, the definition is rather hypothetical. Nevertheless, this definition of probability is in widespread use and is the basis of the frequentist statistics. However, avoiding 'prior probabilities' has its price.

- For many problems no frequency distribution or sample is available at all. So it is not possible to define a probability in these cases, for example:
  - Is Mr X guilty?
  - Does the temperature of the fusion plasma $T \in (1.0, 2.0)10^8$ K?
  - Was Julius Caesar left-handed?
- Even if relative frequencies can be determined, only very rarely is $N \gg 1$ achievable. This is typically the case in:
  - large-scale experiments – only a few dedicated experiments are performed due to limited budget;
  - astrophysics – the number of observations is given by nature (e.g. the number of neutrinos from a supernova).

- The limit $N \to \infty$ cannot be accessed in practice and therefore equation (1.7) has to be considered as a hypothesis which defies experimental validation.
- Relative frequencies provide no clear interpretation for specific individual situations:
  - What meaning can be assigned to a statement like: The probability for Mr X being guilty is 0.05? Does it imply that of 100 clones of Mr X, five would be guilty?
  - The probability for drawing the main prize in a lottery is 10^{-8}. Does this mean that buying 10^8 lottery tickets results in a sure win?

The last two examples also reveal how different probabilities are perceived, depending on the context. In the first situation the probability for Mr X's guilt would be considered to be small, whereas in the second example – despite the very small probability for success – a sufficient number of people happily buy lottery tickets. This is because the very high probability for a small loss (the expense of a lottery ticket) seems overcompensated by the potential gain. The idea of assessing probabilities in the light of the outcome leads to 'decision theory'. Within the framework of this theory outcomes are assigned a value (or loss) and the best decision minimizes the expected loss. This topic is addressed in more detail in Chapter 28 [p. 491].

The statistical approach can address only a limited subset of problems directly, those where relative frequencies are appropriate. In many cases only indirect conclusions can be drawn, and that will be discussed later on.

1.3 Bayesian understanding of 'probability'

Bayesian probability theory is the consequent continuation of the Laplacian approach. It is based on propositions, i.e. statements which are either true or false with nothing in between, like

- S_1: It will rain tomorrow.
- S_2: Rolling a fair perfect die N times will yield face value '3' n times.
- S_3: The time between two radioactive decays lies in the interval $[t, t + dt]$.
- S_4: The variance specified by the manufacturer of the device is wrong.
- S_5: Next time the coin will land with heads up.
- S_6: Theory 'T' correctly describes this phenomenon.
- S_7: The coin in my hand is a cent.

In Bayesian reasoning, the probability $P(S_n|\mathcal{I})$ is a measure for the correctness or truth of proposition S_n and Bayesian probability theory provides a consistent calculus for these probabilities. Roughly speaking, it is the generalization of the propositional calculus to partial truth. In Bayesian probability theory all propositions are on par, no matter whether they describe denumerable random events (S_2), continuous random events (S_3), or even situations where the uncertainty originates solely from missing information (S_7).

Above, we have introduced the notation $P(S_n|\mathcal{I})$, which is actually a conditional probability. It stands for the probability for the proposition S_n, given the 'background information' $\mathcal{I}$, which is also called a conditional complex. The background information uniquely