PART I

FOUNDATIONS

CHAPTER

1 INTRODUCTION

The everyday world around us is a mixture of predictability and chance. The "exact sciences," such as physics or chemistry, are built on the idea that there are definite laws governing observable phenomena, and these sciences have had extraordinary success in elucidating the relevant laws, and in using these laws to describe the world.

Life sciences and social sciences are in a different position. In describing the growth and decline of biological populations, or working out how money makes the world go round, there is always the tantalizingly obvious fact that there is an element of regularity in the phenomena observed, but at the same time nothing is exactly predictable. Economists are willing to say that a financial crisis will happen, but can't say when—and even they will admit they may be wrong. A reliable medical test always has some false positives and negatives, and even something as certain as death comes at a time that can never be determined precisely.

Even in the case of the exact sciences, measured data are only connected with theoretical predictions with a certain margin of error, which in practice can usually be made very small, but cannot be entirely eliminated. And when looked at on the level of individual atoms, all processes become probabilistic, and quantum mechanics provides the appropriate description.

Probability theory is almost unquestioned as the correct way to handle the reality that the information we have about the world is never exact. In some sense, probability theory is the form that logic takes when nothing is absolutely certain. We will therefore start with an outline of the fundamentals of probability theory.

1.1 EVENTS AND THEIR PROBABILITIES

For our purposes, an event is a general concept, which covers ideas such as:

i) There are n bacteria in a sample.

ii) A particle is within a small volume $d^3\boldsymbol{x}$ centered on the position $\boldsymbol{x}$.

https://doi.org/10.1063/9780735423718_001

Elements of Stochastic Methods, Crispin Gardiner

iii) There were n_1 viruses in a cell at time t_1 and n_2 viruses in the cell at time t_2.

These three possibilities exemplify the kinds of situation with which we would want to associate the idea of a probability.

1.1.1 Sets of Events

Mostly, events fall into two categories, those that are specified by sets of integers, and those characterized by sets of continuous variables. The first kind is specified by a vector of integers

$$\boldsymbol{n} = (n_1,\ n_2,\ n_3\ \ldots)\,. \tag{1.1}$$

These are countable events.

The second kind of event is not countable, and is specified by a vector of real numbers

$$\boldsymbol{x} = (x_1,\ x_2,\ x_3\ \ldots)\,. \tag{1.2}$$

More generally, it is useful to consider sets of events, which we can label as A, B, $C\ldots$, etc. For countable events, a set can be specified by listing the events contained in it. The second kind of set can be specified in terms of unions and intersections of volumes ΔV in the space to which $\boldsymbol{x}$ belongs.

1.1.2 Notations

We will use the notations

$$\Omega \ : \text{ The set of all events,} \tag{1.3}$$

$$\varnothing \ : \text{ The set of no events,} \tag{1.4}$$

$$\{\boldsymbol{a}\} \ : \text{ The set containing only the event } \boldsymbol{a}. \tag{1.5}$$

1.1.3 Probabilities

Probability is most simply defined in terms of sets of events, A, within the space of all events of the kind we wish to consider. We introduce the quantity $P(A)$ as the probability that an arbitrary event is contained in A.

a) Axioms: The probability must satisfy the following probability axioms for all sets:

i) *Positivity*:

$$P(A) \geqslant 0. \tag{1.6}$$

This is a formalization of the intuitive belief that a probability is proportional to the number of times that something happens, which is clearly either positive or zero.

ii) *Completeness*:

$$P(\Omega) = 1. \tag{1.7}$$

This is the expression of the fact that every event is certain to be contained within Ω.

iii) *Mutually Exclusive Events*: If A_i $(i = 1,2,3,\ldots)$ is a countable (but possibly infinite) collection of non-overlapping sets, that is

$$A_i \bigcap A_j = \varnothing \qquad \text{for all} \quad i \neq j, \tag{1.8}$$

then

$$P\Big(\bigcup_i A_i\Big) = \sum_i P(A_i)\,. \tag{1.9}$$

The condition that the sets are non-overlapping is the formal statement that events in the various sets are mutually exclusive, and the axiom states that their probabilities simply add.

The number of sets must be countable because of the existence of sets labeled by a continuous index, for example $\boldsymbol{x}$, the position in space. The probability of a molecule being in the set whose only element is $\boldsymbol{x}$ is zero, but the probability of it being in a region R of finite volume, or an infinitesimal region such as $d^3\boldsymbol{x}$, is non-zero. The regions R and $d^3\boldsymbol{x}$ can both be expressed as a union of sets of the form $\{\boldsymbol{x}\}$—but not a *countable* union. Thus axiom iii) is not applicable and the probability of being in R or $d^3\boldsymbol{x}$ cannot be expressed as the sum of the probabilities of being in $\{\boldsymbol{x}\}$.

b) Corollaries: Two further facts follow from the axioms.

i) If $\bar{A}$ is the complement of A, i.e., the set of all events not contained in A, then $A \cup \bar{A} = \Omega$, and hence from ii)

$$P(\bar{A}) = 1 - P(A)\,. \tag{1.10}$$

ii) As a special case, since $\varnothing = \bar{\Omega}$, it follows that

$$P(\varnothing) = 0\,. \tag{1.11}$$

1.1.4 Relating Probability to the Real World

The probabilities that we have introduced cannot be directly and rigorously related to the real world. The classic example of tossing dice illustrates this immediately. Intuitively, we expect each of the values 1 to 6 will have the same probability of occurring. Obviously, it is possible to weight the dice to favor a particular number, or perhaps to use some sleight of hand to toss the dice to achieve the same end. We exclude this—the dice must be constructed "fairly" and tossed "fairly." This means that we must construct and toss the dice so that the outcome is uncertain, and equally likely to happen. The reasoning is, of course, circular.

By eliminating what we now think of as intuitive ideas and axiomatizing probability, *Kolmogorov* [1.1] cleared the road for a rigorous development of mathematical probability. His insight was to recognize that the definition of what we mean by probability in the real world is not a mathematical question, and that the above axioms are both in correspondence with reality, and sufficient to formulate probability as a branch of mathematics.

The simplest way of looking at axiomatic probability is as a formal method of manipulating probabilities using the axioms. In order to apply the theory, the probability space must be defined *and* the probability measure P assigned. These are *a priori probabilities*, which are assigned on grounds appropriate to the system under study. The task of applying probability is:

i) To assume some set of a priori probabilities that seem reasonable and to deduce results from this and from the structure of the probability space.

ii) To measure experimental results with some apparatus that is constructed to measure quantities in accordance with these a priori probabilities.

1.2 JOINT AND CONDITIONAL PROBABILITIES

1.2.1 Joint Probabilities

We explained in Sec. 1.1.3 how the occurrence of mutually exclusive events is related to the concept of non-intersecting sets. We now consider the concept $P(A \cap B)$, where the intersection $A \cap B$ is non-empty. An event $\boldsymbol{a}$ within A will only be within $A \cap B$ if it is also within B as well, hence

$$P(A \cap B) = P\{(\boldsymbol{a} \in A) \text{ and } (\boldsymbol{a} \in B)\}, \tag{1.12}$$

and $P(A \cap B)$ is called the *joint probability* that the event $\boldsymbol{a}$ is contained in both classes—that is, that both the event $\boldsymbol{a} \in A$ and the event $\boldsymbol{a} \in B$ occur.

1.2.2 Relationship Between Joint Probabilities of Different Orders

Suppose that we have a collection of sets B_i such that

$$B_i \cap B_j = \varnothing, \qquad \bigcup_i B_i = \Omega, \tag{1.13}$$

so that the sets divide up the space Ω into non-overlapping subsets.

Then

$$\bigcup_i (A \cap B_i) = A \cap \Big(\bigcup_i B_i\Big) = A \cap \Omega = A. \tag{1.14}$$

Using now the probability axiom iii), we see that $A \cap B_i$ are a countable collection of non-overlapping sets, and therefore satisfy the conditions on the A_i used there. Hence

$$\sum_i P(A \cap B_i) = P\big(\bigcup_i (A \cap B_i)\big) = P(A), \tag{1.15}$$

and more generally,

$$\sum_i P(A_i \cap B_j \cap C_k \ldots) = P(B_j \cap C_k \cap \ldots). \tag{1.16}$$

Thus, summing over all mutually exclusive possibilities of B in the joint probability eliminates that variable.

1.2.3 Conditional Probabilities

We need to define *conditional probabilities*, which are defined only on the collection of all sets contained in B.

We define the conditional probability as

$$P(A|B) = P(A \cap B)/P(B), \tag{1.17}$$

and this satisfies our intuitive conception that the conditional probability that $\boldsymbol{a} \in A$ (given that we know $\boldsymbol{a} \in B$) is given by dividing the probability of joint occurrence by the probability ($\boldsymbol{a} \in B$).

Using the definition(1.17) of the conditional probability, and the result (1.15), it follows that

$$\sum_i P(A|B_i)P(B_i) = P(A). \tag{1.18}$$

This kind of result has very significant consequences in the development of the theory of Markov processes, which will be considered in detail in Chap. 4.

1.2.4 Independence

We need a probabilistic way of specifying what we mean by independent events. Two sets of events A and B should represent independent sets of events if the specification that a particular event is contained in B has no influence on the probability of that event belonging to A. Thus, the conditional probability $P(A|B)$ should be independent of B, and hence

$$P(A \cap B) = P(A)P(B)\,. \tag{1.19}$$

In the case of several events, we need a somewhat stronger specification.

a) Definition of Independent Events: Events are considered to be independent if their joint probabilities factorize. More precisely, the events $\left(a_i \in A_i,\ i = 1,2,\ldots,n\right)$ will be considered to be independent if for any subset $(i_1, i_2, \ldots, i_k)$ of the set $(1,2,\ldots,n)$,

$$P(A_{i_1} \cap A_{i_2} \ldots\ A_{i_k}) = P(A_{i_1})P(A_{i_2})\ \ldots P(A_{i_k})\,. \tag{1.20}$$

b) All Possible Factorizations are Necessary: It is important to require factorization for all possible combinations, as in (1.20).

1.3 PROBABILITY NOTATIONS

The set-theoretic formulation of probability used in the previous sections is powerful, and convenient for discussing general issues. When used in practice, it is more natural to use notations that are more specific to the actual situations under consideration. In particular, we will introduce the ideas of probability distributions and probability densities, as well as the idea of random variables.

1.3.1 Probability Distribution Function

For applications, the probability of occurrence of a single value $\boldsymbol{n}$ is the most convenient quantity to use. This corresponds to considering sets of only one event, such as $\{\boldsymbol{n}\}$, and it is convenient to use the notation

$$P(\boldsymbol{n}) \ \equiv\ P\left(\{\boldsymbol{n}\}\right). \tag{1.21}$$

$P(\boldsymbol{n})$—the probability of occurrence of the value $\boldsymbol{n}$—is then called the probability distribution function.

1.3.2 Probability Density

For a probability space with members taking on a continuous range of values $\boldsymbol{x}$ in a space of r dimensions, we consider the probability associated with a set of points in an infinitesimal volume $d^r\boldsymbol{x}$ around the value $\boldsymbol{x}$, and write this in the form

$$P(\boldsymbol{x}, d^r\boldsymbol{x}) \equiv p(\boldsymbol{x})\, d^r\boldsymbol{x}. \tag{1.22}$$

This defines $p(\boldsymbol{x})$ as the probability density function for this system.

1.3.3 Random Variables

The idea of a random variable is a way of talking about a probability space and the associated probabilities using a single symbol. For example, if we have a probability distribution $P(n)$, where $n = 0, 1, 2, \ldots$, we can define the random variable N as a quantity that takes on the values n with probability $P(n)$. This notation means that any function $f(N)$ of N is itself a random variable, which takes on the values $f(n)$ with probability $P(N)$. For example, N might mean the number of molecules in a small volume ΔV. This is a quantity which we do not know exactly, but which can be reasonably described as taking on the value n with probability $P(n)$.

1.3.4 Independent Random Variables

Random variables $N_1, N_2, N_3, \ldots$, will be said to be independent random variables, if their joint probability distribution function factorizes as in Sec. 1.2.4, that is,

$$P(n_1, n_2, \ldots, n_i, \ldots) = P_1(n_1) P_2(n_2) \ldots P_i(n_i) \ldots. \tag{1.23}$$

For all sets of the form $A_i = (x \text{ such that } a_i \leqslant x \leqslant b_i)$, the events $N_1 \in A_1, N_2 \in A_2, N_3 \in A_3, \ldots$ are independent events. This will mean that all values of the N_i are assumed independently of those of the remaining N_i.

1.4 MEAN VALUES OF RANDOM VARIABLES

1.4.1 Definitions

a) Countable Events: The mean value (or expectation) of a discrete random variable $\boldsymbol{N}$ is given by

$$\langle N \rangle = \sum_{\boldsymbol{n}} n P(n). \tag{1.24}$$

b) Notation for the Mean Value: The notation $\langle N \rangle$ for the expectation used in this book is a physicist's notation. The more common mathematical notation is $E(N)$.

c) Events Described by a Probability Density: In this case, the mean value of a random variable is given by integration

$$\langle \boldsymbol{X} \rangle = \int \boldsymbol{x}\, p(\boldsymbol{x})\, d^r \boldsymbol{x}. \tag{1.25}$$

d) The Variance: The variance $\mathrm{var}[X]$ of the random variable X is given by

$$\mathrm{var}[X] \equiv \big(\sigma[X]\big)^2 \equiv \big\langle \big(X - \langle X \rangle\big)^2 \big\rangle. \tag{1.26}$$

As is well known $\mathrm{var}[X]$, or its square root the standard deviation $\sigma[X]$, is a measure of the degree to which the values of X deviate from the mean value $\langle X \rangle$.

1.4.2 Some History

The now almost universal acceptance of the mean as a representative of the "true value" of a random variable took some time to develop. In his very accessible article, *Stahl* [1.2] notes that it was *Galileo* [1.3] who first considered the properties of random errors inherent in the observations of celestial phenomena, but he did not come to any conclusion as to how a "true" or "best" value corresponding to a set of observations should be estimated. Only at the beginning of the 19th century did the mean become accepted as the most practical measure of the "true value" of a random variable.

1.4.3 The Law of Large Numbers

Let us consider taking a finite number M of samples x_i of the random variable X. We intuitively expect that as M becomes very large, the average

$$\bar{x}_M \equiv \frac{1}{M} \sum_i x_i \tag{1.27}$$

approaches the mean of the random variable $\langle X \rangle$. *Under the condition that the mean and variance exist*, this can be proved. This result, called the law of large numbers, establishes the mean as the preferred estimator of a random variable under these conditions.

The law of large numbers is proved by showing that the variance of the average $\bar{x}_M$ approaches zero as the number of observations M becomes very large. This is done by constructing the random variable corresponding to the average. We do this by considering the M measurements to be

independent samples of the same probability distribution. This effectively constructs a set of independent random variables X_i, all with the same probability distribution. The X_i all have the same mean and variance, which we can write as $\langle X \rangle$ and $\mathrm{var}[X]$.

From these we construct the random variable

$$\bar{X}_M \equiv \frac{1}{M}\sum_{i=1}^{\infty} X_i. \tag{1.28}$$

The mean and variance of $\bar{X}_M$ are of course given by the standard results

$$\langle \bar{X}_M \rangle = \langle X \rangle, \tag{1.29}$$

$$\mathrm{var}\left[\bar{X}_M\right] = \frac{\mathrm{var}[X]}{\sqrt{M}} \longrightarrow 0 \quad \text{as} \quad M \longrightarrow \infty. \tag{1.30}$$

This means that in the limit of large M, the only observable value of the average $\bar{X}_M$ is the mean $\langle X \rangle$ of the random variable X.

This is the law of large numbers—a result that relates the abstract probability concepts to reality.

1.4.4 Applicability of the Law of Large Numbers

It is important to note that the validity of the law of large numbers requires that the variance does exist, and that this condition is not always satisfied. In practice, one would expect that for a very wide range of measurable quantities, the relevant random variable would have a finite range. For example, the ages of members of a human population can be expected to be confined to a finite range of about 130 years at most. In such a case the variance must exist.

1.4.5 Heavy-Tailed Distributions

Situations in which the variance does not exist are not only possible, but in fact are quite important. They are characterized by a slow falloff of the probability density as $|x| \to \infty$—such a distribution has come to be named a heavy-tailed distribution. These are treated in detail in Chap. 3.

As an example, consider the Cauchy distribution, whose probability density is given by

$$p_{\mathrm{Cauchy}}(x) \equiv \frac{1}{\pi}\frac{a}{x^2 + a^2}\,. \tag{1.31}$$

This is a well-defined probability density, and is correctly normalized to 1. However, it is clear that $\int_{-\infty}^{\infty} x^2\, p_{\mathrm{Cauchy}}(x)\, dx$ diverges, so that the variance cannot exist.

Even the mean, which by symmetry one might expect to be zero, can only be defined as the principal value integral

$$\langle X \rangle_{\mathrm{Cauchy}} = \lim_{z\to\infty}\int_{-z}^{z} x\, p_{\mathrm{Cauchy}}(x)\, dx. \tag{1.32}$$

We will discuss the interpretation of the Cauchy distribution in more detail in Sec. 3.4.1.

The Cauchy distribution provides a very accurate description of the frequency distribution of photons emitted during a transition between atomic energy levels, where it is normally called the Lorentzian distribution. In practice, spectral measurements are done by measuring the frequency distribution with a spectrometer. Instead of a mean value and a variance, the distribution is characterized by the position of the maximum, and the full width at half maximum (FWHM), which for the distribution (1.31), has the value $2a$.

1.4.6 Moments, Correlations, and Covariances

The moments $\langle X^n \rangle$ are often seen as quantities by which a probability distribution can be characterized. The mean and variance involve the first two moments, and provide the most elementary way of characterizing a probability distribution. To fully characterize a probability distribution requires the knowledge of all moments. However, because a probability distribution must always vanish as $x \to \pm\infty$, the higher moments tell us only about the properties of unlikely large values of X.

a) Existence of Moments: There is no requirement that the moments of any order actually exist. This is demonstrated by the example of the Cauchy distribution (1.31) as above, and by other heavy-tailed distributions.

b) Several Random Variables: In the case of several variables, we define the covariance matrix as

$$\langle X_i, X_j \rangle \equiv \big\langle (X_i - \langle X_i \rangle)(X_j - \langle X_j \rangle) \big\rangle \equiv \langle X_i X_j \rangle - \langle X_i \rangle \langle X_j \rangle \,. \tag{1.33}$$

Obviously,

$$\langle X_i, X_i \rangle = \mathrm{var}\,[X_i] \,. \tag{1.34}$$

If the variables are independent *in pairs*, the covariance matrix is diagonal.

1.5 THE CHARACTERISTIC FUNCTION

If $\boldsymbol{s}$ is the vector $(s_1, s_2, \ldots, s_n)$, and $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ is a vector of random variables, then the characteristic function (or moment generating function) is defined by the Fourier transform

$$\phi(\boldsymbol{s}) \equiv \langle \exp(\mathrm{i}\boldsymbol{s}\cdot\boldsymbol{X}) \rangle = \int d\boldsymbol{x}\; p(\boldsymbol{x}) \exp(\mathrm{i}\boldsymbol{s}\cdot\boldsymbol{x}) \,. \tag{1.35}$$

Because of the Fourier inversion formula

$$p(\boldsymbol{x}) = (2\pi)^{-n} \int d\boldsymbol{s}\, \phi(\boldsymbol{s}) \exp(-\mathrm{i}\boldsymbol{x}\cdot\boldsymbol{s}), \tag{1.36}$$

$\phi(\boldsymbol{s})$ determines $p(\boldsymbol{x})$ with probability 1. Hence, the characteristic function does truly *characterize* the probability density.

1.5.1 Properties of the Characteristic Function

The characteristic function has the following properties:

i) The most important property is that $\phi(\boldsymbol{s})$ exists for any probability density function. It therefore provides a much more useful tool than the moments, which as we have seen do not always exist.

ii) $\phi(\boldsymbol{s})$ is a uniformly continuous function of its arguments for all finite real $\boldsymbol{s}$.

iii) $\phi(\boldsymbol{0}) = 1$.

iv) $\left|\phi(\boldsymbol{s})\right| \leqslant 1$.

v) If the *moments* $\left\langle \prod_i X_i^{m_i} \right\rangle$ exist, then they are given in terms of the characteristic function by the derivatives:

$$\left\langle \prod_i X_i^{m_i} \right\rangle = \left[\prod_i \left(-\mathrm{i}\frac{\partial}{\partial s_i} \right)^{m_i} \phi(\boldsymbol{s}) \right]_{\boldsymbol{s}=\boldsymbol{0}} . \tag{1.37}$$

Conversely, when moments do not exist, the corresponding derivative of the characteristic function does not exist at $\boldsymbol{s} = \boldsymbol{0}$. For example, the characteristic function of the Cauchy distribution (1.31) is $\exp(-a|s|)$, for which no derivatives exist at $s = 0$.

vi) A sequence of probability densities converges to a limiting probability density if and only if the corresponding characteristic functions converge to the corresponding characteristic function of the limiting probability density.

vii) *Independent random variables* $X_1, X_2 \ldots X_n$: The definition of independence in Sec. 1.2.4 shows that the set of variables $X_1, X_2 \ldots X_n$ are independent if and only if

$$p(x_1, x_2, \ldots, x_n) = p_1(x_1)\, p_2(x_2) \ldots p_n(x_n), \tag{1.38}$$

in which case,

$$\phi(s_1, s_2, \ldots s_n) = \phi_1(s_1)\phi_2(s_2) \ldots \phi_n(s_n). \tag{1.39}$$

viii) *Sum of independent random variables*: If $X_1, X_2, \ldots$ are independent random variables, u_i are constants, and if

$$Y = \sum_{i=1}^{n} (u_i X_i + v_i), \tag{1.40}$$

and the characteristic function of Y is

$$\phi_y(s) = \langle \exp(\mathrm{i}sY) \rangle \,, \tag{1.41}$$

then

$$\phi_y(s) = \prod_{i=1}^{n} \mathrm{e}^{\mathrm{i}v_i s} \phi_i(u_i s) \,. \tag{1.42}$$

1.5.2 Role and Significance of the Characteristic Function

The characteristic function plays an important role, which arises from the convergence property vi). This allows us to perform limiting processes on the characteristic function rather than the probability distribution itself, and often makes proofs easier. As well as this, the straightforward derivation of the moments by (1.37) makes any determination of the characteristic function directly relevant to measurable quantities.

REFERENCES FOR CHAPTER 1

[1.1] A. N. Kolmogorov, *Foundations of the Theory of Probability* (Chelsea, New York, 1950; the German original appeared in 1933). 1-4

[1.2] S. Stahl, *The evolution of the normal distribution*, Math. Mag. **79**, 96 (2006). 1-8

[1.3] G. Galilei, *Dialogue Concerning the Two Chief World Systems—Ptolemaic & Copernican* (Univ. California Press, Berkeley, 1967). 1-8