2 Probability

2.1 Modeling Measurements: A Need for Probability

Gambling aficionados are accustomed to the notion that when they roll a pair of dice repeatedly, they obtain, each time, very different number combinations, and unless the dice have been heavily tampered with, it is impossible to control or predict which faces will roll up. The same may be said of the balls in a lottery machine: it is not possible to actually predict which balls will be drawn. This seems rather obvious. Is it not? Well, actually no! The classical mechanics that govern these processes is wholly deterministic. Given enough information about the initial kinematic conditions of a roll as well as the properties of the dice and the table they are rolled on, it should be possible to calculate, at least in principle, which face the dice will actually land on. This is perhaps less obvious, but it is true. Phenomena such as a roll of dice or the breakdown of a window by an impact are ruled by deterministic laws of physics and should thus be predictable. The problem, of course, is that they are immensely complicated phenomena involving a large succession of events. For instance, a proper calculation of the trajectory of a set of dice would require knowledge of their exact speed and orientation when they leave the player's hand. One would also need to account for friction, the air pressure and temperature, the exact elasticity of the dice and all components of the table on which they roll, etc. And because dice can bounce several times against each other and on the table, one should have to follow their complete trajectory, accounting for whatever imperfections the dice or the table might feature. This is a rather formidable task that is unlikely to ever be accomplished, even with modern supercomputers. Effectively, if dice are thrown with enough vigor, they should bounce and roll so many times as to make the calculation practically impossible. For all intents and purposes, the roll of fair dice is truly a random phenomenon: its outcome is unpredictable and all faces have an equal probability of rolling up.

Scientific experiments are obviously not a form of gambling, but the many physicochemical processes involved in measurements have much in common with a dice roll. Typical measurements involve a large succession of macroscopic and microscopic processes that randomly alter their outcome. Effectively, repeated measurements of a given physical quantity (e.g., the position, momentum, or energy of an object) also yield different values that seemingly fluctuate and adopt a random pattern, which, at best, clusters near the actual value of the observable of interest.

The fluctuations stem in part from the technique used to carry out the measurements. For instance, ten people who measure the length of, say, a table with a tape measure, will

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

report slightly different values, even though they might use the same tape measure. The differences have to do with the way these different people position the tape measure next to the table, how they read the tape measure, whether they keep it extended without stretching it, and so on. Measurement accuracy may also be limited by the physical process used to carry out the measurement as well as the observable measured. For example, scatterings and energy losses in a detector randomly affect the momentum of the particles measured in a magnetic spectrometer. If a given particle is kicked this or that way, its direction and energy are slightly altered, and the momentum determined based on the particle's trajectory is slightly off from the actual value.

The bottom line is that measured values of physical observables are likely to vary from measurement to measurement. Additionally, inspection of the values obtained in a sequence of measurements will reveal that the specific value obtained in a given measurement cannot be predicted. The measured values appear as a sequence of seemingly random numbers. Yet a table has a length, and measurements of this length yield values that cluster around a typical value that one expects should be representative of the actual length of the table. Likewise, repeated measurements of particle positions, momenta, and so on shall yield values that vary from measurement to measurement but should cluster around the actual values of these observables.

Throughout this book, we adopt the view that the outcome of measurements of scientific observables is, for all intents and purposes, a random process. The outcomes of a given measurement, or a succession of measurements, shall be considered random variables. But randomness does not imply all values of an observable are equally probable: a reasonably well designed and carefully carried out experiment shall yield values that cluster near and about the actual value of the observable. Whoever has carried out a succession of measurements of a well-defined observable according to a specific measurement protocol can in fact attest that such clustering occurs. We will assume that it is possible, at least in principle, to formulate a probabilistic model of the measurement process and the clustering of its outcomes. In other words, although it is impossible to predict the value of a specific measurement with absolute certainty, it should be possible to formulate a model that provides the probability of any given outcome or the distribution of values obtained after several measurements. Although each specific outcome is random, the probability model should describe, overall, the probability of any given outcome according to a specific mathematical function called the probability distribution.

Evidently, a model of the measurement process shall be as good as the efforts put into understanding it as well as the prior knowledge available about the process and the apparatus used to carry out the experiment. The job of a statistician shall thus be to find and apply statistical methods that enable a trustworthy characterization or **inference** of measured observables and scientific models used to describe phenomena of interest, even though the measurement model may not be completely accurate. A difficulty arises that statistical inference is contingent on the notion of probability and how this notion is applied to the description and characterization of experimental results.

Two main paradigms, known as frequentist and Bayesian, are commonly used to define and interpret the notion of probability. The frequentist paradigm assumes measurement fluctuations are determined by a parent distribution representing the relative frequency of

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

values or ranges of values of an observable. The parent distribution is a priori unknown but can be determined, at least in principle, in the limit of an infinite number of measurements. The Bayesian paradigm dispenses with the need for an infinite number of measurements by shifting the discussion into a hypothesis space in which all components of the measurement process, including the probability model of the measurement and its parameters, are considered as hypotheses, each endowed with a degree of plausibility, that is, a probability.

It is the purpose of §2.2 of this chapter to motivate and introduce the two interpretations of probability these paradigms are based on. A discussion of the pros and cons of the two interpretations is initiated in §2.3 but will continue throughout the first part of this book. The remainder of this chapter discuss mathematical concepts pertaining to and used in both paradigms. Section 2.4 introduces Bayes' theorem, the notions of inference, as well as the concepts of sample and hypothesis space, while §2.5 defines discrete and continuous random variables, probability distribution, probability density distributions, as well as cumulative distributions and densities. Sections 2.6 and 2.7 next introduce functions of random variables and techniques for the characterization of distributions. Multivariate distributions and their moments are discussed in §2.8 and §2.9, respectively. Section 2.10 introduces the notions of characteristic function, moment-generating function, and examples of their application, including a proof of the very important central limit theorem. With these tools in hand, we proceed to introduces notions of measurement errors and random walk processes in §2.11 and §2.12, respectively. The chapter ends, in §2.13, with the definition of cumulants, which have broad utility in probability and statistics, and constitute, in particular, an essential component of correlation and flow analyses conducted in the field of high-energy heavy-ion collisions.

2.2 Foundation and Definitions

The notion of probability can be intuitively introduced on the basis of the relative frequency of the phenomena of interest. For instance, to establish that a cubic die has been tampered with, it suffices, in principle, to roll it a very large number of times, and count how many times each face rolls up. The relative frequency of each face, f_i, with $i = 1, \ldots, 6$ is the number of times, N_i, each face rolls up divided by the total number of rolls, N_{Tot}.

$$f_i = \frac{N_i}{N_{\text{Tot}}} \tag{2.1}$$

The frequencies f_i provide an indication of the likelihood, that is, the **probability**, of rolling any face i in subsequent rolls of this same die. A die is considered fair if all faces have the same probability, that is, the same frequency.

The notion of frequency as a probability also applies in virtually all forms of experimental measurement. For example, repeated measurements of the universal gravitational constant, G (Big G), are expected to yield values in the neighborhood of the actual value of the constant. The number of times values are observed in a specific ranges $[G, G + dG]$, relative to the total number of observations, can then be used to estimate the probability

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

of observing G in that range. It is thus natural to establish a connection between notions of probability and relative frequency. This leads to the **frequentist interpretation** of the notion of probability.

The frequentist interpretation of probability is most concerned with the outcome of measurements or observations. The observations, either discrete or continuous numbers, are elements of sets, known as **sample space**, defining the domain of observable values. One can thus naturally and formally introduce the notion of probability based on values associated with elements or subsets of the sample space. Such a definition, introduced in 1933 by Kolmogorov, is presented in §2.2.1. The frequentist interpretation becomes problematic if and when dealing with systems involving phenomena that cannot be observed more than once (e.g., the collapse of the Tacoma Narrows bridge, a star exploding in a supernova, or the Big Bang), or when trying to express the plausibility or truthfulness of statements about the world. How indeed does one quantify the probability that a statement about dark matter or the Higgs boson is correct if it does not involve numerical values, whether discrete or continuous (e.g., dark matter does not consist of known baryons, or there exists only one Higgs boson)? Obviously such statements are the result of inference based on prior knowledge of the world and the outcome of experimental measurements with specific measured data. One then needs a robust and well-defined method to translate one's certainty about prior knowledge and experimental results into statements about the world. One must also be able to combine such statements according to the rules of logic. This then creates the need to extend logic (or the calculus of predicates) to include the notion of plausibility or likelihood. Such an extension was developed largely by E. T. Jaynes and his collaborators in the 1970s. We briefly present the basic tenets and rules of the foundation of probability as logic in §§ 2.2.3 and 2.2.4.

2.2.1 Probability Based on Set Theory

The concept of probability embodies the notion of randomness, in other words, the fact that one cannot predict with complete certainty the outcome, or value, of a particular measurement. It may be formally defined in the context of **set theory**. Indeed, measurement outcomes can be viewed as members of a set. We call **sample space**, denoted **S**, the set consisting of elements corresponding to actual and possible outcomes of a measurement. The set **S** can be divided into subsets **A**, **B**, **C**, and so on. A subset **A** of **S**, noted $\mathbf{A} \subseteq \mathbf{S}$, may be empty and contain no elements. It may alternatively contain one, few, or all elements of **S**. It is then possible to assign **A** with a real number, noted $p(A)$. This number is called the **probability of A** provided it obeys the following three axioms:

$$p(A) \geq 0 \text{ for } \mathbf{A} \subseteq \mathbf{S}, \tag{2.2}$$

$$p(A \cup B) = p(A) + p(B) \text{ for } \mathbf{A} \cap \mathbf{B} = 0, \tag{2.3}$$

$$p(S) = 1. \tag{2.4}$$

The first axiom stipulates that every subset **A** belonging to **S** has a probability $p(A)$ larger than or equal to zero. The second axiom states that the probability assigned to the union

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

$\mathbf{A} \cup \mathbf{B}$ of two disjoint subsets $\mathbf{A}$ and $\mathbf{B}$ (i.e., $\mathbf{A} \cap \mathbf{B} = 0$) is the sum of their probabilities $p(A)$ and $p(B)$. The third axiom states that the probability assigned to $\mathbf{S}$ is unity.

Introducing the notation $\overline{\mathbf{A}}$ to mean the complement of $\mathbf{A}$ in $\mathbf{S}$, such that $\mathbf{A} \cup \overline{\mathbf{A}} = \mathbf{S}$, one can demonstrate the following properties (see Problem 2.1):

$$p(\bar{A}) = 1 - p(A), \tag{2.5}$$

$$p(\bar{A} \cup A) = p(S) = 1, \tag{2.6}$$

$$0 \leq p(A) \leq 1, \tag{2.7}$$

$$\text{if } \mathbf{A} \subset \mathbf{B}, \text{ then } \; p(A) \leq p(B), \tag{2.8}$$

$$\text{if } \mathbf{A} \text{ is a null subset, then } p(A) = 0. \tag{2.9}$$

The first expression indicates that the probability of the complement of $\mathbf{A}$ is equal to 1 minus the probability of $\mathbf{A}$ itself. The second expression indicates that the probability of the union of $\mathbf{A}$ and its complement, which can be identified as $\mathbf{S}$, has a probability equal to one. The third expression states that the probability of $\mathbf{A}$ can take any value in the range [0, 1], including the values 0 and 1. The last expression states that the probability of an empty subset is null.

Consider a "measurement" of an observable (variable) X that can take any specific value or group of values corresponding to elements of the set $\mathbf{S}$. Further consider that a measurement of X might yield any values in the set. This variable is then considered a **random variable**, and the axioms (2.2–2.4) define the probability of measuring X within a given subset $\mathbf{A}$. Note that X may represent a single value, a range, or combinations of values, as we will illustrate in the text that follows. Additionally, if values of X are restricted to the set of integers, such as in counting experiments (e.g., number of people recovering from an illness thanks to a medication or the number of particles produced in a nucleus–nucleus collision), X is said to be a **discrete random variable**, whereas if X belongs to a subset or the entire set of real numbers $\mathbf{R}$, it constitutes a **continuous random variable**.

2.2.2 Conditional Probability and Statistical Independence

Let us now consider a sample space $\mathbf{S}$, with subsets $\mathbf{A}$ and $\mathbf{B}$ such that $p(B) \neq 0$. One then defines the **conditional probability**, noted $p(A|B)$, as the probability of $\mathbf{A}$ given $\mathbf{B}$:

$$p(A|B) = \frac{p(A \cap B)}{p(B)}. \tag{2.10}$$

This corresponds to the probability of observing the random variable X within $\mathbf{A}$ when it is also within $\mathbf{B}$. It is relatively straightforward to show that the notion of conditional probability satisfies the axioms of probability introduced previously (see Problem 2.2). In fact, the probability $p(A)$ can itself be viewed as a conditional probability $p(A) = p(A|S)$ since $p(S) = 1$ by construction.

Two subsets $\mathbf{A}$ and $\mathbf{B}$, and the measurement outcomes they represent, are said to be **independent** if they satisfy the condition

$$p(A \cap B) = p(A)p(B). \tag{2.11}$$

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

This condition means that the probability that X is a member of **A** and **B simultaneously** is equal to the product of the probabilities of X being in **A** and **B** independently. This enables the evaluation of the conditional probability $p(A|B)$:

$$p(A|B) = \frac{p(A \cap B)}{p(B)} = \frac{p(A)p(B)}{p(B)} = p(A) \quad \text{(statistical independence)}. \tag{2.12}$$

Consequently, when **A** and **B** are statistically independent, the conditional probability of **A** given **B** is equal to the probability of **A** itself, in other words, the probability of **A** does not depend on **B**. Likewise, the conditional probability of **B** given **A** is equal to the probability of **B** itself:

$$p(B|A) = \frac{p(A \cap B)}{p(A)} = \frac{p(A)p(B)}{p(A)} = p(B) \quad \text{(statistical independence)}. \tag{2.13}$$

It is quite important to realize that the notion of **statistical independence** or **independent subsets**, $p(A \cap B) = p(A)p(B)$, differs from that of **disjoint subsets**, $\mathbf{A} \cap \mathbf{B} = 0$. Indeed, if the intersection $\mathbf{A} \cap \mathbf{B}$ is empty, it is not possible for an element x (a measurement outcome) to be simultaneously part of **A** and **B**; the probability of observing such an element (measurement value) is thus null, and therefore different in general from $p(A)p(B)$.

Clearly, the conditional probabilities $p(A|B)$ and $p(B|A)$ are not independent. To establish their relation, first consider that by definition of the conditional probability, one has

$$p(A|B) = \frac{p(A \cap B)}{p(B)}, \tag{2.14}$$

and similarly,

$$p(B|A) = \frac{p(B \cap A)}{p(A)}. \tag{2.15}$$

Given the commutativity of the intersection of two sets, $\mathbf{B} \cap \mathbf{A} = \mathbf{A} \cap \mathbf{B}$, one finds that the two conditional probabilities are related as follows:

$$p(A \cap B) = p(B|A)p(A) = p(A|B)p(B), \tag{2.16}$$

provided neither $p(A)$ nor $p(B)$ is null. This implies it is possible to calculate the conditional probability $p(B|A)$ as follows:

$$p(B|A) = \frac{p(A|B)p(B)}{p(A)}. \tag{2.17}$$

This expression, known as **Bayes' theorem**,[1] is a fundamental relationship in probability theory and finds use in a wide range of practical applications, as we shall discuss in this and subsequent chapters of this book.

It is useful to further explore the properties of conditional probabilities by partitioning the set **S** into finitely many subsets, $\mathbf{A}_i$ with $i = 1, \ldots, n$. Consider, for instance, n subsets $\mathbf{A}_i$, whose union is by construction equal to **S**, in other words, such that $\mathbf{S} = \bigcup_i \mathbf{A}_i$. Assume

[1] Thomas Bayes (1702–1761) was an English mathematician and Presbyterian minister known for the development of the mathematical theorem that bears his name.

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

further that none of these subsets are null, $p(A_i) \neq 0$, and that they are all disjoint, $\mathbf{A}_i \cap \mathbf{A}_j = 0$ for $i \neq j$. Then it is possible to show (see Problem 2.3) that

$$p(B) = \sum_i p(B|A_i)p(A_i). \tag{2.18}$$

This result is known as the **law of total probability**.

Combining the law of total probability with Bayes' theorem, one finds

$$p(A|B) = \frac{p(B|A)p(A)}{\sum_i p(B|A_i)p(A_i)}, \tag{2.19}$$

which is particularly useful in the estimation of the probability of a specific hypothesis given measured results. We present an example of an application of Bayes' theorem and the law of total probability in the following subsection.

Example: A Problem about Problematic Parts

Statement of the Problem

10P100Bad, a famous auto-parts supplier, conducted a study of the parts it produces in a year. The study revealed 90% of the produced parts are within specifications while the remainder 10% are not. Stricken with revenue losses, the company decided to hide this fact and sell both the good and defective parts. A client, the struggling car company Don'tFoolMe, suspected there was a problem with the parts and designed its own test to determine whether the components it bought from 10P100Bad were within tolerances. Unfortunately, their test was not very precise. They estimated that their test had a probability of 1% to identify a component as nondefective even though it was, and a probability of 0.5% to reject a nondefective part. Calculate the probability a bad part would not be rejected by their test and used in the construction of a car.

Solution

Let us establish what we know from the statement of the problem. First, we know that the prior probability, $p(D)$, a part is defective is 10%:

$$p(D) = 0.1, \tag{2.20}$$

$$p(\bar{D}) = 0.9. \tag{2.21}$$

We also know the probability, $p(A|D)$, of accepting a defective part is 1%,

$$p(A|D) = 0.01, \tag{2.22}$$

$$p(\bar{A}|D) = 0.99, \tag{2.23}$$

and the probability, $p(\bar{A}|\bar{D})$, of rejecting a nondefective part is 0.5%. One thus writes

$$p(A|\bar{D}) = 0.995, \tag{2.24}$$

$$p(\bar{A}|\bar{D}) = 0.005. \tag{2.25}$$

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

We now seek the probability, $p(D|A)$, that a part accepted in the fabrication of the cars is defective. By virtue of the total probability theorem, one writes

$$\begin{aligned} p(D|A) &= \frac{p(A|D)p(D)}{p(A|D)p(D) + p(A|\bar{D})p(\bar{D})}, \\ &= \frac{0.01 \times 0.1}{0.01 \times 0.1 + 0.995 \times 0.9}, \\ &= \frac{0.001}{0.001 + 0.8955} = 0.0011. \end{aligned} \tag{2.26}$$

One then concludes the car maker Don'tFoolMe has a probability of 0.1% of integrating defective parts in its fleet. That sounds like a financial disaster in the making. . .

2.2.3 A Need for Probability as an Extension of Logic

The notion of probability based on a sample space as a set of numbers representing the outcome of measurements is rather restrictive. Indeed, as already stated, one would also like to quantify the plausibility of general statements about the world or specific phenomena. For instance, modern cosmologists are concerned with the notion that the expansion of the universe (e.g., the fact that all observable galaxies recede from one another at speeds that grow proportionally to the distance that separate them) might accelerate over time. One would then like to use existing data to quantify the plausibility of statements such as the expansion of the universe has been constant through times, or the expansion has accelerated during the last n billion years, and so forth. Indeed, one would like to express a probability for either statements given the state or prior knowledge about the universe and measured data (e.g., based on type Ia supernovae). In all scientific endeavors, one is interested in stating/quantifying the plausibility of statements made about the world or specific phenomena. The problem, of course, is that one typically deals with limited data and that all data involve finite errors.

The scientific process, applied to a specific inquiry of a specific system, involves predictive statements about measurements based on one or several models of the system (e.g., space is warped by the presence of massive objects such as stars). Models can then be used to make predictions concerning phenomena occurring in the realm of the system (e.g., the warp of space around a star was used by A. Einstein to predict the deflection of star light by the Sun observable during an eclipse of the Sun and a sizable contribution to the precession of planet Mercury in its orbit around the Sun). Measurements of observables of interest (e.g., a shift in the apparent position of stars near the line of sight of the eclipse, and the measured rate of precession of Mercury's orbit) can then be used to infer, after due statistical analysis of the measured values and their errors, whether the model predictions are correct. It then becomes natural to enlarge (or replace) the notion of sample space with the notion of hypothesis space consisting of (logical) statements about the phenomenon in question (or the world in general). Probability may then be viewed as an extension of logic where propositions (predicates) are ascribed a plausibility, likelihood, or degree of belief.

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

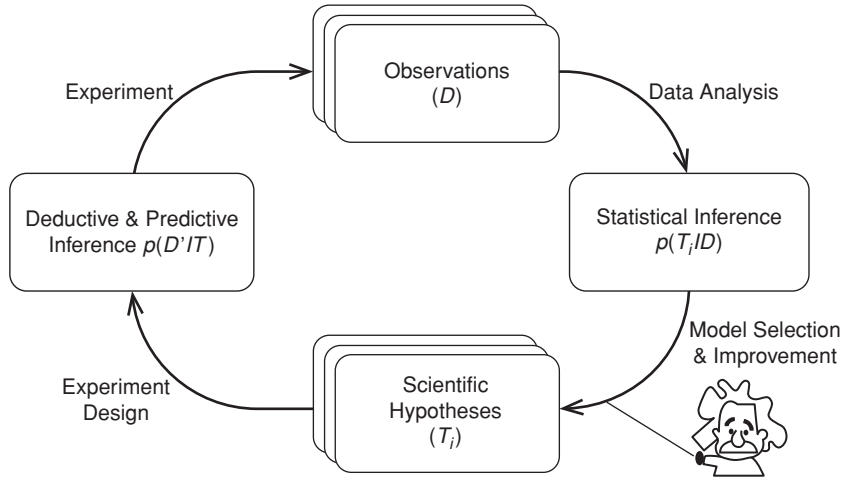

Fig. 2.1 Schematic summary of the scientific method as a cyclical process involving stages of modeling, deductive (predictive) inference, measurements, statistical analyses, and inference.

A proper foundation of this approach was progressively developed by several authors including C. Shannon [172], H. Jeffreys [118, 119], and R. T. Cox [69], arguably culminating with the work of Jaynes published posthumously in 2003 [117]. Several articles and books have extended Jaynes' work.

An extension of the notion of probability to include logic is attractive because one can use the formally well-established rules of logic to combine logic statements and reason based on these predicates in a sound and robust manner. Of course, the point of associating probabilities to logical statements is that one can then express the probability of some statements being true if logically derived from (or entailed by) prior statements when the veracity of such statements is itself not perfectly well established. The probability of a statement derived from several others thus depends on the individual probabilities of these statements.

Probability as logic is also particularly useful because it enables a scientific discourse based on logic while accounting formally and robustly for the limited amount of information one may have about a specific phenomenon (or perhaps the universe as a whole). The scientific method, which nominally involves the formulation of hypotheses and the realization of measurements designed to test these hypotheses, may then be formulated as a cycle, illustrated in Figure 2.1, involving predictive inference based on models, deployment of one or several experiments whose results are analyzed statistically (statistics) and used to carry out formal tests of the hypotheses or models (statistical inference), and eventually to decide which of the models has the highest probability, that is, is most consistent with the measured data. This may then be followed by theoretical work leading to additional or tweaked hypotheses and models, which then form the basis for new experiments or measurements. The cycle repeats, eventually leading to significant scientific advances.

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

Physical models of nature are well covered in courses on classical mechanics, electrodynamics, quantum field theory, and so forth. But such courses typically ignore the details of the predictive process leading to the formulation of experiments, the statistical analysis of the data, and the statistical inference involved in transforming observations (measurements) into conclusions about the veracity or appropriateness of the models. One of the purposes of this book is to fill this gap by providing a relatively detailed presentation of the methods of predictive inference, data analysis techniques (including techniques used to correct for instrumental effects, e.g., biases and defects), parameter estimation, hypothesis testing, and statistical inference.

2.2.4 Probability as an Extension of Logic

Extending logic to include the notions of plausibility and probability is not a trivial matter. Much like the generalization of geometry to non-Euclidean space, much freedom appears to exist, at least at the outset, in the manner in which this can be done. Cox [69] and Jaynes [117] formulated such an extension starting from three **desiderata**, that is, three sets of properties or attributes a theory of probability based on logic should satisfy. The desiderata and the foundational theoretical structure are here stated without much of a discussion or proof. Readers interested in digging into the foundations of the theory should consult the book by Jaynes [117] or the more recent book by Gregory [97].

Cox and Jaynes have reasoned that an extension of logic including probabilities should satisfy the following three desiderata:

1. The degree of plausibility of statements is represented by real numbers.
2. The measure of plausibility must behave rationally: as new information supporting the truth of a particular statement (predicate) is obtained, its plausibility must increase continuously and monotonically.
3. A theory of probability as logic must be consistent:
   a. There must be structural consistency: if a conclusion can be reasoned along many paths, all paths (based on the same information) must yield the same result, that is, the same value of plausibility.
   b. It should be possible to account for all information relevant to a particular problem in a quantitative manner. In other words, it should be possible to assign a degree of plausibility to prior information and account for it in the reasoning process.
   c. Equivalent states of knowledge must be represented by equivalent values of plausibility.

Jaynes demonstrated that these desiderata entail two rules, a sum rule and a product rule, that provide a foundation for a probability theory based on logic. We here state these two rules without demonstration. Derivations of these rules may be found in refs. [97, 117].

The sum rule is written

$$\text{Sum Rule}: \quad p(A|B) + p(\overline{A}|B) = 1 \tag{2.27}$$

The notation $p(X)$, stated p of X, represents the probability (plausibility) of the predicate X, which may consist of any properly constructed combinations of simpler predicates.

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

The symbols A and B represent two such logic predicates (rather than subsets of a sample space), that is, statements about the world or a particular phenomenon. The vertical bar is used to indicate that one considers the probability of the predicate on the left, given or assuming that the predicate on the right is true. For instance, the notation $p(A|B)$ corresponds to the probability that the statement A is true when the statement B is known to be true. A short horizontal line over a letter is here used to indicate the negation of the statement: $\overline{A}$ (commonly stated non-A) represents the logical negation of the statement A. The sum rule encodes the rather obvious and sensible notion that the sum of the probability of A being true given B and the probability of its negation $\overline{A}$ (also given B) is equal to unity and therefore exhausts all possibilities.

The product rule concerns conditional probabilities and is written

$$\text{Product Rule}: \quad p(A, B|C) = p(B|A, C)p(A|C) \tag{2.28}$$

The comma notation A, B expresses a logical conjunction, that is, a logic "AND" between the propositions A and B. The statement A, B can be true if and only if both A and B are true. The notation $p(A, B|C)$ thus expresses the probability of A, B being true when C is known to be true, whereas $p(B|A, C)$ corresponds to the probability of B being true when A, C is true, that is, when it is known that both A and C are true. The product rule tells us that the probability of A, B being true given C is known to be true is equal to the product of the probability of B being true when A and C are known to be true, and the probability of A being true when C is known to be true.

Clearly, since the conjunction operation commutes, that is, $A, B = B, A$, the product rule may also be written

$$p(A, B|C) = p(B, A|C) = p(A|B, C)p(B|C) \tag{2.29}$$

We will see in the text below that the product rules entails Bayes' theorem but let us first consider the relation between the sum and product rules to the axioms of probability stated in §2.10.

An attentive reader will have noted that the two rules are not mere accidents but were essentially designed to be consistent with the axioms of probability based on set theory. Consider, for illustrative purposes, two statements A and B as follows:

- A: Random variable X is found in the subset **A** of the sample space **S**.
- B: Random variable X is found in the subset **B** of the sample space **S**.

Let us first consider a case where **B** is the complement of **A** in **S**, that is, $\mathbf{B} = \overline{\mathbf{A}}$. The sum rule applied to the statement A tells us

$$p(A|S) + p(\overline{A}|S) = 1 \tag{2.30}$$

Substituting B in lieu of $\overline{A}$, one has

$$p(A|S) + p(B|S) = 1 \tag{2.31}$$

In the language of set theory (§2.2.1), this may be written

$$p(A \cup B) = p(A) + p(B) = 1 \tag{2.32}$$

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

given $\mathbf{A} \cap \mathbf{B} = 0$ by virtue of the definition $\mathbf{B} = \overline{\mathbf{A}}$. The sum rule, Eq. (2.27), thus embodies the axioms given by Eqs. (2.2–2.4).

The product rule similarly includes and extends the conditional probability definition given by Eq. (2.10): let C state that the random variable X is found within $\mathbf{S}$. Since $\mathbf{S}$ corresponds, by construction, to the entire sample space spanned by X, the statement C is always true by construction. A logical AND between C and some arbitrary statement D is thus equal to D ($D, C = D$). The proposition C can thus be omitted. The product rule

$$p(A, B|C) = p(A|B, C)p(B|C) \tag{2.33}$$

may then be written

$$p(A, B) = p(A|B)p(B). \tag{2.34}$$

For the statement A, B to be true, the variable X must be found simultaneously in A and B. This is possible only if the intersection of these two subsets, $\mathbf{A} \cap \mathbf{B}$, is nonempty. One then finds that the conditional probability $p(A|B)$ is given by

$$p(A|B) = \frac{p(A \cap B)}{p(B)}, \tag{2.35}$$

which is precisely the definition of conditional probability given by Eq. (2.10).

Let us now return to the expression of the product rule given by Eq. (2.28). As stated earlier, thanks to commutativity of the *AND* operation, one may write

$$p(A, B|C) = p(B, A|C). \tag{2.36}$$

By virtue of the product rule, this expression may then be written

$$p(A|B, C)p(B|C) = p(B|A, C)p(A|C), \tag{2.37}$$

which corresponds to Bayes' theorem:

$$p(A|B, C) = \frac{p(B|A, C)p(A|C)}{p(B|C)} \tag{2.38}$$

Indeed, following the same line of arguments as earlier in this paragraph, that is, identifying C as stating that the variable is found in the sample space $\mathbf{S}$, one recovers Eq. (2.17):

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}. \tag{2.39}$$

We note in closing this section that the sum rule may also be written

$$\text{Extended Sum Rule}: \; p(A + B|C) = p(A|C) + p(B|C) - p(A, B)|C). \tag{2.40}$$

The plus sign is here used to represent a logical disjunction, that is, a logical OR operation. The expression $p(A + B|C)$ thus corresponds to the probability of A or B being true, given that C is known to be true. It is relatively straightforward to show that this expression is consistent with the simpler sum rule (Eq. 2.27) as well as the axioms (2.2–2.4).

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

2.3 Frequentist and Bayesian Interpretations of Probabilities

The definitions based on set theory and probability as logic discussed in the previous sections naturally lead to two distinct interpretations of the notion of probability. In one, the so-called **frequentist interpretation**, a probability is regarded as a limiting value of the relative frequency of an outcome (either a discrete number or a range of values) when the number of trials becomes infinite. In the other, often referred to as **subjective probability**, the notion of probability expresses the plausibility of specific statements, which can be interpreted as degree of belief said statements are true. It is also commonly referred as **Bayesian interpretation** of probability.

Scientists, particularly physicists, have long made use of the frequentist interpretation and developed a great many tools to assign errors to measurements, and conduct statistical tests of scientific hypotheses. Studies of phenomena where the notion of limiting frequencies does not readily apply have prompted many scientists to pay more attention and embrace the Bayesian interpretation in their experimental studies and toward the inference of conclusions derived on their experiments. The Bayesian interpretation is now used in a growing number of scientific applications.

2.3.1 Frequentist Interpretation

The **frequentist interpretation** derives its name from the fact that a probability can often be regarded as a **limiting relative frequency**. In this context, the elements of a set **S** correspond to the possible outcomes of a measurement considered to be repeatable an **arbitrary large** number of times. A subset **A** of **S**, commonly referred to as an **event**, amounts to a set of possible outcomes of the measurement or observation. A particular event is said to occur whenever a measurement yields a member of a given subset (e.g., the subset **A**). A subset consisting of one element denotes a single elementary outcome. The probability of **A** thus corresponds to the fraction of all events yielding this particular outcome in the limit when measurements are hypothetically repeated infinitely many times:

$$p(\mathrm{A}) = \lim_{n\to\infty} \frac{\text{number of occurrences of } \mathbf{A}}{n}. \tag{2.41}$$

The probability of the occurrence of nonelementary outcomes may be determined from the probability of individual outcomes, consistent with the axioms expressed in Eqs. (2.2–2.4) because, by construction, the fraction of occurrences is always greater than zero and less than or equal to unity. The frequentist interpretation of probability forms the basis of the branch of mathematics known as **classical statistics**, also known as **classical inference** and **frequentist statistics**. All tasks, techniques, and methods based on the frequentist interpretation of probability are also said to form or be part of the **frequentist inference paradigm**.

Critics of the frequentist approach argue that given an infinite number of measurements is clearly not possible, the limit $n \to \infty$ cannot be achieved or verified in practice. The probability $p(A)$ of a set of outcomes, A, consequently cannot be determined with perfect

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

precision. Effectively, one must assume that a particular measurement may be represented by a specific parent probability distribution. A frequentist statistician must then establish, on the basis of a finite number of measurements, whether a particular distribution or model properly describes the measurement(s) at hand. This may fortunately be accomplished on the basis of statistical tests discussed in §6.4. One may then compare several models and tests, based on finite data samples, determine which model is best compatible with the measured data.

In spite of the aforementioned conceptual difficulty, the frequentist interpretation is used routinely in science texts on probability and statistics, and by scientists in their analyses. It is typically considered appropriate and sufficient whenever one deals with scientific observations that can be repeated many times. It is, however, somewhat problematic whenever a measurement or event (e.g., the Big Bang, a supernova, a volcanic eruption, or the collapse of a bridge) cannot be repeated. It also makes it impossible to directly and explicitly integrate scientific hypotheses in the probabilistic discourse.

2.3.2 Subjective Interpretation

The **subjective interpretation of probability**, also called **Bayesian interpretation of probability**, is used increasingly in the physical sciences and many other scientific fields. It can be formulated based on both set theory and probability as logic, but we will argue, throughout this text, that a formulation based on probability as logic is far more interesting, convenient, and powerful.

Within the foundation of probability based on set theory, an event is regarded as a statement that the observable X is an element of the subset $\mathbf{A}$. The quantity $p(A)$ may then be interpreted as the degree of belief the statement $\mathbf{A}$ might be true:

$$p(A) = \text{degree of belief that the statement } \mathbf{A} \text{ is true.} \tag{2.42}$$

In this context, the Bayesian interpretation assumes it is possible to construct the sample space $\mathbf{S}$ in terms of elementary hypotheses that are mutually exclusive, in other words, implying that only one statement is actually true. A set consisting of multiple such disjoint subsets is therefore true if any one of the subsets it contains is true. And one then has $p(S) = 1$.

Jaynes' definition of probability as an extension of logic readily extends the realm of the probability discourse. Indeed, dealing with predicates rather than sets and subsets, it becomes naturally possible to discuss the probability of statements about any world entities that can be expressed within the calculus of predicate. This implies, in particular, that the probabilistic discourse is no longer confined to the outcome of measurements and observations (being members of sets) but can be augmented to include statements about models, scientific hypotheses, and so forth. In this context, the quantity $p(A)$ may then be interpreted as the degree of belief the proposition A might be true:

$$p(A) = \text{degree of belief that the proposition } A \text{ is true,} \tag{2.43}$$

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

where the proposition A is not restricted to statements about measurement outcomes but can include statements about models, scientific hypotheses, and so forth.

The concept of subjective probability is closely related to Bayes' theorem and forms the basis of **Bayesian statistics** and **Bayesian inference**. It also forms the basis of the **Bayesian inference paradigm**.

In this context, one can then consider probabilities of the form $p(B|A)$, where A expresses a specific scientific hypothesis (e.g., a statement about a model or a model parameter), while B might represent the hypothesis that a specific experiment will yield a specific outcome (i.e., a specific discrete value or a continuous value in specific range). The conditional probability $p(B|A)$ then represents the degree of belief that B is observed given a hypothesis A is true. As such, the Bayesian inference paradigm provides a convenient framework, discussed in great detail in Chapter 7, to gauge the merits of one or competing models (or theories) relating to a specific measurement or set of measurements.

Given a certain theory, T, one might assign a certain **prior probability**, $p(T)$, that this theory is a valid model of the world (or set of experimental results). The probability, $p(D|T)$, called the **likelihood**, then provides an estimate of the degree to which measured data, D, can be expected based on the theory T. The conditional probability $p(T|D)$ thus provides a **posterior probability** that the theory, T, is true, conditioned by the data (measurements). According to Bayes' theorem, this posterior may be written

$$p(T|D) \propto p(D|T) \times p(T). \tag{2.44}$$

In this context, data are considered as facts and thus taken as true.[2] Bayes' theorem then enables the evaluation of the probability $p(T|D)$ that the theory T might be true, given the data. In other words, the merits of the theoretical hypothesis T can be gauged and evaluated based on the available data. This leads to the notion of hypothesis testing, which is first discussed in the context of the frequentist paradigm in Chapter 6 and more directly and naturally within the Bayesian paradigms in Chapter 7.

Given a dataset D, the merits of different theories or hypotheses, $T_1, T_2, \ldots$ can in principle be compared. Ideally, a particular theory T_i might emerge to have a posterior probability much larger than the others, $p(T_i|D) \gg p(T_j|D)$ for $j \neq i$, and would then become the favored theory. In practice, it is often the case that several competing models or hypotheses yield relatively weak and similar posterior probabilities. The available data are then considered insufficient to discriminate between the models.

It is fair to note that Bayesian statistics does not provide, ab initio, any particular method to determine the prior probability, $p(T)$. In the absence of prior inferences based on other theories, models, or data, it might be set to unity. The likelihood probability, $p(D|T)$, then provides the sole basis for the evaluation of $p(T|D)$. In other situations, there could be

[2] One should bear in mind, however, that measured values might need substantial corrections to be fully representative of the observable of interest. This implies that the probability model accounting for a specific measurement should include a proper probabilistic description of relevant instrumental effects or that raw measurement values can be "corrected" to account for such instrumental effects. This important topic is discussed in Chapter 12.

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

older data that enables an evaluation of the prior $p(T)$ before the experiment is conducted. The "new" data can then be seen as improving the knowledge about T. Quite obviously, the value of $p(T|D)$ is subject to the prior hypothesis as well as the data. This then leads to a framework that is subjective, hence the notion of subjective interpretation of probability. Although this might be seen as a weakness, one should stress that once a prior $p(T)$ and the likelihood $p(D|T)$ are determined, Bayes' theorem unambiguously provides an estimate of the posterior probability $p(T|D)$. In a scientific context, this provides for a mechanism to submit models and theories to strict and constraining tests of validity. Examples of such tests are presented in Chapter 7.

> Unfortunately, Bayesian statistics is often considered in opposition to classical (frequentist) statistics. In fact, some problems discussed within the frequentist and Bayesian paradigm yield contrasting and incompatible solutions. Hard-core frequentists argue that the notion of degree of belief in a prior hypothesis leads to arbitrary posteriors and thus reject the Bayesian paradigm altogether. Some Bayesian statisticians argue that the definition of probability in terms of a limit, Eq. (2.41), is itself artificial or arbitrary, and thus reject the frequentist paradigm. Can there be a common ground?

It may be argued that Bayesian statistics in fact includes the frequentist interpretation as a special case, and as such provides a broader and more comprehensive context for data analysis. The formulation of probability as logic discussed in §2.2.4 naturally embodies the Bayesian interpretation of probability. Indeed, probability defined as an extension of logic deals with predicates or statements about the world (or a particular phenomenon), and assigns a certain degree of plausibility to these predicates. Predicates are thus viewed as elements of a hypothesis space rather than a simple set of numerical values. It is then possible, as we already argued, to consider more general and elaborate problems of inference. We will come back to this idea in more detail in Chapter 7. This said, it should also be clear that predicates considered in a particular analysis may also be formulated solely on the basis of sets of values, and the corresponding probabilities of these values can then be viewed as limiting frequencies, that is, frequencies that would be observed should an infinite number of observations be made. For instance, it is reasonable to consider that the outcome of a measurement will yield a certain element of $\mathbf{S}$ a certain fraction of the time. A prior, $p(A)$, may thus be regarded as the degree of belief that a certain probability distribution dictates the outcome of a measurement. The conditional probability $p(B|A)$ then provides the degree of belief that the given probability distribution yields an outcome B within $\mathbf{S}$. The subjective interpretation thus effectively encompasses the relative frequentist interpretation if one admits the implicit proviso that $p(A) = 1$. A subjective interpretation may, however, also be associated with cases in which the concept of frequency is not readily or meaningfully applicable. For instance, while the notion that a certain quantity X lies within a specific interval can be determined in both interpretations, the determination of confidence intervals with the frequentist interpretation assumes it is reasonable to use a specific parent probability distribution to carry out the calculation of the interval. In effect, this assumes one has a reasonably high degree of belief that a specific probability distribution is a proper representation of the outcome of an experiment. Effectively, the prior, which corresponds to the probability that a specific probability determines the outcome of

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

a measurement, is assumed to have maximal probability. The frequentist interpretation can thus indeed be viewed as "special case" of the subjective interpretation.

2.4 Bayes' Theorem and Inference

Whether working within the frequentist interpretation or the Bayesian interpretation of probability, Bayes' theorem is ideally suited toward statistical inference analyses, that is, analyses where one wishes to establish the optimal value of model parameters, estimates of their errors, or which of many competing hypotheses has the highest probability of asserting the truth about a particular system or phenomenon. Although frequentist inference can and will be considered in this context, it is far more convenient to introduce the concept of inference within the Bayesian paradigm using the notion of probability as an extension of logic because generic statements about scientific hypotheses (i.e., models, model parameters, etc.) can be evaluated in a single formal and robust mathematical setting where the prior plausibility of hypotheses as well as data are considered.

2.4.1 Basic Concepts of Bayesian Inference

Let H_i, with $i = 1, \ldots, n$, represent a set of n propositions asserting the truth of competing hypotheses. Given the very nature of the scientific process, these hypotheses are formulated out of a particular context. Let us represent relevant statements from this context (also known as prior information) as I. We will additionally represent measured data in terms of a proposition D. For inference purposes, Bayes' theorem may then be written

$$p(H_i|D, I) = \frac{p(D|H_i, I)p(H_i|I)}{p(D|I)}. \tag{2.45}$$

The quantity $p(D|H_i, I)$ represents the probability of observing the data D if both H_i and I are true. It is commonly called likelihood of the data D based on the hypothesis H_i, or simply likelihood function, and noted $\mathbf{L}(H_i)$. The quantities $p(H_i|I)$ and $p(H_i|D, I)$ represent the prior and posterior probability of the hypothesis H_i. The probability $p(H_i|I)$ is based solely on prior knowledge whereas $p(H_i|D, I)$ includes both the prior knowledge and the new knowledge provided by the measurement D. The denominator, $p(D|I)$, is seemingly more cryptic but it corresponds to the probability of obtaining the data D given the prior information available on the system or phenomenon. Although it may be difficult to assess this probability directly, note that it can be computed in terms of the law of total probability

$$p(D|I) = \sum_i p(D|H_i, I)p(H_i|I) \tag{2.46}$$

where the sum is taken over all hypotheses that can be formulated about the system. All in all, this factor provides a normalization factor that ensures that the sum over the probability

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

of all hypotheses, given the data and prior information, is equal to unity:

$$\sum_i p(H_i|D, I) = 1. \tag{2.47}$$

2.4.2 Hypothesis vs. Sample Space

Within the frequentist approach, one is focused on the outcome of measurements and techniques mostly to utilize these measurements to extract information about a phenomenon or system. Measured observables may be discrete (e.g., number of particles observed in a specific proton–proton collision) or continuous (e.g., the momenta of produced particles). Observed values, collectively called **sample**, may then be viewed as elements of either a discrete set (i.e., a subset of **Z**, the set of integers) or a continuous set (i.e., a subset of **R**, the set of real numbers). The measured values are thus considered random outcomes, or random variables, either discrete or continuous, from a **parent population** known as a **sample space**.

The Bayesian approach shifts the focus toward statements about the data and hypotheses or models used to described the data. The goal is indeed to use the measured data to establish the plausibility (or degree of belief) of various hypotheses or statements formulated about a system (phenomenon) and the data it produces. Hypotheses may concern various characterizations of the data, model parameters, or even a model as a whole. They may be formulated either in terms of discrete statements (e.g., dark matter exists; there is only one Higgs boson; etc.) or in the form of continuous statements (e.g., the Hubble constant lies in the range $[H_0, H_0 + dH_0]$; the mass of the Higgs boson is in the range $[M, M + dM]$, etc.). The Bayesian approach thus enlarges, so to speak, the sample space associated with the outcome of measurement observables to include a space of hypotheses or model statements that can be made about a system both before and after the measurement is conducted. It is then concerned with assigning degrees of belief, or plausibility, to each of these hypotheses or model statements.

Strictly speaking, hypotheses are not random variables, but specific statements about a phenomenon or reality at large. Indeed, dark matter either exists or does not, but it is not a random phenomenon. Likewise, physical quantities such as the speed of light or Planck constant have specific values and thus cannot be legitimately regarded as random variables. The true (and precise) values of the observables factually remain unknown, however, so the Bayesian notion of degree of belief that the value of a physical quantity might lie in a given interval thus makes good sense. Yet, measurements involve a number of effects that may effectively smear or seemingly randomize observed values. A Bayesian statistician must then account for the measurement outcomes with a probability model of the measurement process. One concludes that while an observable of interest might not be random, measured instances of the observable will invariably appear random. Consequently, insofar as probabilities are regarded as degrees of belief, there is no philosophical difficulty or contradiction in considering prior probabilities that an observable X might lie within a range $[x_0, x_0 + dx]$ while measurement instances have a probability $p(x|x_0)$, determined by the

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

measurement process, to be found in the range $[x, x + dx]$. Effectively, we conclude that both x_0 and x can be treated as random variables.

2.5 Definition of Probability Distribution and Probability Density

Whether one adheres to the frequentist or Bayesian interpretation of probabilities, one is faced with either discrete or continuous variables. With finitely many discrete values, it is obviously possible to assign a (finite) probability to each value separately but if the number of discrete values is infinite, or if the variables are continuous, one must introduce the notion of probability density. We discuss basic features and properties of discrete variable first, in §2.5.1, and consider continuous variables next, in §2.5.2.

2.5.1 Discrete Observables and Probability Distribution Functions

Consider, for instance, a game of dice in which players throw two cubic dice at a time on a mat. The faces of the dice are labeled with numbers ranging from 1 to 6. The game may then involve betting on the sum of the dice values rolled in a given throw. Clearly, the sum of dice rolled takes only a finite number of discrete values from 2 to 12 and is thus a **discrete outcome**. There is only one way to get a sum of 2 or 12, but several ways to roll a 6 or 7. The outcome of the bet should then be decided based on the probability of a given roll.

One should remark, once again, that a roll of dice is nominally a phenomenon that can be described in terms of deterministic laws of physics. Indeed, given specific initial conditions (i.e., the position, orientation, translational and rotational speed of the dice), one could in principle predict the outcome of a roll provided the elastic properties of the dice and the table on which they roll are well known. In practice, the properties of the dice and table are not so well known, and measuring the initial conditions of a roll with sufficient precision is rather tricky. In essence, not enough is known about the system (the two dice and mat) to enable an accurate calculation of the outcome of a roll, that is, on which face the dice will stop rolling. The outcome of a dice roll thus appears unpredictable and the sum of the top faces may then be regarded as a random variable. Given the geometrical symmetry of a die, it is natural to assume all faces are equally likely. Within the frequentist interpretation, and for a fair die, one expects that all six faces should have the same frequency after rolls have been repeated a very large number of times, while in the Bayesian interpretation, one may ab initio express the belief (or plausibility) that the perfect symmetry of a die implies each face has a probability of $1/6$ of rolling up.

More generally, one may be concerned with the determination of the probabilities of values taken by one or several discrete random variables. For example, a marketer might be concerned with the number of people showing up at a special public event based on ads published in newspapers or played on radio stations, whereas an astronomer might be interested in counting how many supernovae explosions were detected in a specific night with a powerful telescope. In these and other discrete systems, as for a roll of dice, one

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

assumes the systems are not sufficiently well known (either by virtue of their macroscopic complexity or their inherent nondeterministic character) to predict a specific outcome with certainty, and one must then assess either a frequency (frequentist approach) or degree of plausibility (belief) that specific values might be observed.

In the case of a measurement of one discrete random variable, n, the sample space consists of all (integer) values, or combinations of values, the variable can take. Assuming the number of such values is finite, one is then concerned with the probability, $p(n)$, of each element individually. In a roll of a pair of dice, for instance, one might want to know the probability of rolling a 7. The quantity $p(n)$ is thus a function that represents how the probability of values of n is distributed across the sample space $\mathbf{S}$ and is known as the **Probability Distribution Function**, or PDF.

By virtue of the third axiom, Eq. (2.4), the sum of the probabilities of all outcomes must be unity. The sum of the probabilities $p(n)$ must thus satisfy the condition:

$$\sum_{n \in \mathbf{S}} p(n) = 1. \tag{2.48}$$

2.5.2 Continuous Observables and Probability Density Functions

Obviously, there are also cases in which measured observables can take continuous values. Examples of continuous variables include the temperature of the atmosphere at sunset, the barometric pressure during a storm, the strength of the electric and magnetic fields produced by an antenna, or the momentum of particles produced by nuclear collisions, and so on. One may be interested in studying how such quantities vary with time, position, or other variables. Alternatively, one might be interested in the very precise determination of "constants" of nature, such as the speed of light in vacuum, the lifetime of the ^{14}C radioisotope, or the cross section of a particular nuclear reaction. Within the context of the Bayesian approach, one may also be interested in considering continuous hypotheses. This is the case, for instance, when a particular model parameter or observable cannot be observed directly but must be inferred from one or several other measurements. One can then formulate continuous hypotheses stating that a continuous observable O lies with a given range $[O, O + \Delta O]$.[3]

While a physical quantity is known (or assumed) to have a single and unique value, repeated measurements would yield continuous values that fluctuate, seemingly arbitrarily, from measurement to measurement. One is thus faced with **continuous random variables**, which are either elements of a continuous sample space (frequentist approach) or a continuous hypothesis space (Bayesian approach). Either way, a space of continuous (random) variables, or combination of random variables, is obviously infinite. It is thus not meaningful to talk about the probability of a specific value. One is instead concerned with the probability of measuring values in specific finite intervals (e.g., $[O, O + \Delta O]$). However,

[3] Note that the same letter is here used to represent both a logical proposition and the observable it is concerned with.

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

in the limit of vanishingly small intervals, $\Delta O \to 0$, one can introduce the notion of **probability density**.

Let us consider an experiment whose outcome consists of a single continuous observable X. The sample space **S** associated with this measurement may thus consists of a subset of **R**, the set of real numbers. Given continuous subsets of **R** have infinite cardinality, it is not meaningful to consider the probability of a single value, x. One can, however, consider the probability that such an observed value x will be found within the infinitesimal interval $[x, x + dx]$. We shall here assume this probability exists and can be evaluated with a function $f(x)$ known as **Probability Density Function**, hereafter noted, PDF:

$$\text{Probability to observe } X \text{ in } [x, x + dx] = f(x)\, dx. \tag{2.49}$$

Note 1: In general, statisticians use capital letters (e.g., X, Y, W, etc.) to denote or identify the name of observables (observable quantities, variables), while lowercase letters (e.g., x, y, w) are used to label specific instances of these variables. For instance, the variable identifying the position of a particle might be defined as X while a specific measurement (i.e., a specific instance) of the observable would be typically written as x. In physics, however, lowercase and uppercase letters are typically used to denote different quantities or observables (although perhaps related). The uppercase/lowercase convention used by statisticians may then become difficult to apply. We thus (mostly) adhere to the convention in Part I when introducing generic and foundational concepts and relinquish the convention in Parts II and III of the book when discussing physics concepts.

In the frequentist interpretation, $f(x)\, dx$ corresponds to the fraction of times the value x is found in the interval $[x, x + dx]$ in the limit that the total number of measurements is infinitely large, while in the Bayesian interpretation, this quantity provides the degree of belief an observed value x might be in that range, without any particular assumption as to whether the experiment can actually be repeated. Additionally, note that in this context, a continuous variable X could also represent a statement or hypothesis about a parameter of a model used to describe the system or phenomenon. It is thus legitimate to consider probability densities in hypothesis space as well as in sample space.

By virtue of the third axiom (Eq. 2.4), a PDF must be normalized such that the probability of any outcome is unity. The sum of the probabilities of all outcomes is thus the integral of the function $f(x)$ over the entire sample space **S** (or hypothesis space as the case may be) spanned by the observable X:

$$\int_{\mathbf{S}} f(x)\, dx = 1. \tag{2.50}$$

It is important to reemphasize that the notion of probability density function applies to both measurement outcomes and continuous model hypotheses. In fact, much of the discussion that follows in this and following chapters about probability densities is applicable to continuous variables in both sample and hypothesis spaces without much regard as to whether they are discussed in the context of either interpretation.

Note 2: Throughout this text, we use the abbreviation PDF and the notations $f(x)$ and $p(x)$ for both probability distribution functions and probability density functions. In

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

general, it shall be clear from the context whether one is considering discrete or continuous variables and, correspondingly, probability distributions or probability densities.

2.5.3 Cumulative Distribution and Density Functions

It is useful to introduce the notions of **Cumulative Distribution Function** and **Cumulative Density Function**, both hereafter denoted CDF. Given a PDF $f(x)$, the CDF $F(x)$, may be formally defined as a cumulative (sometimes called **running**) sum or integral of the function $f(x)$. For a discrete probability distribution, one has

$$F_n = \sum_{i=0}^{n} f(x_i). \tag{2.51}$$

whereas for a density one has

$$F(x) = \int_{-\infty}^{x} f(x')\,dx'. \tag{2.52}$$

Obviously, the function $F(x)$ amounts to the probability a random variable X takes a value smaller or equal to x. One can thus alternatively first define $F(x)$ as the probability of obtaining an outcome less than or equal to x and evaluate the PDF, $f(x)$, as a derivative of F:

$$f(x) = \frac{dF(x)}{dx}. \tag{2.53}$$

The two definitions are equivalent if the distribution $F(x)$ is well behaved, that is, if it is everywhere differentiable. This can be formalized mathematically, but the notion of "everywhere differentiable" will be sufficient for the remainder of this text.

The concept of cumulative distribution is illustrated in Figure 2.2, which displays the cumulative integral, Eq. (2.52), of the uniform distribution (top panel) and the **standard normal distribution** (bottom panel). The definitions of these distributions and their properties are formally introduced in §§3.4 and 3.9. Note that since both distributions are symmetric relative to their **median**,[4] indicated with a vertical dash line, $F(x)$ takes a value of 0.5 at that point. This means there is a probability of 50% that the value of x will be smaller than $x = 0$. Additionally, note that in the limit of $x \to \infty$, the cumulative function $F(x)$ converges to unity; in other words, the probability of observing any value of x is unity. This is indeed guaranteed by the normalization of $f(x)$ defined by Eq. (2.50).

It is useful to introduce an alternative notation for PDFs. Given a PDF $f(x)$ expresses the probability of occurrences, or events, being observed in the interval $[x, x + dx]$, one may also write

$$f(x) \equiv \frac{1}{N}\frac{dN}{dx}, \tag{2.54}$$

[4] The notion of median is formally defined in §2.7.

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

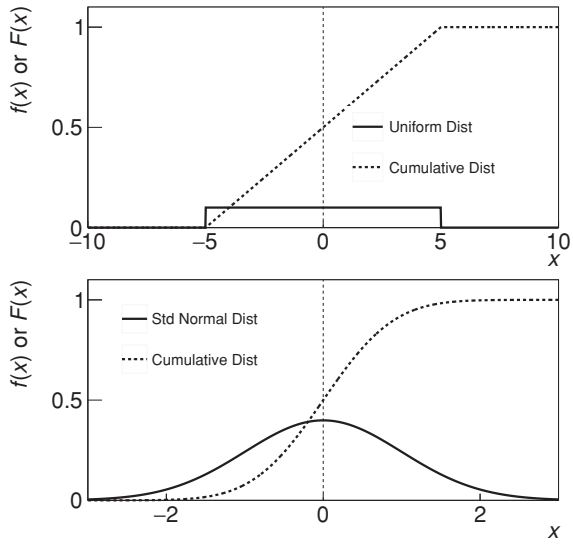

Fig. 2.2 Illustration of the concept of cumulative distribution for a uniform distribution (top) and (bottom) a standard normal (Gaussian) distribution.

where N represents the integral of the function dN/dx over its domain which corresponds to the sample (or hypothesis) space of x. One gets

$$\int_{-\infty}^{\infty} \frac{1}{N}\frac{dN}{dx}dx = \frac{1}{N}\int dN = \frac{1}{N}N = 1, \tag{2.55}$$

which satisfies the normalization condition given by Eq. (2.50). The function dN/dx thus represents a density of counts or events in the interval $[x, x + dx]$, and dividing this density by its integral yields the probability density $f(x)$ with proper normalization.

The function notation, $f(x)$, has the advantage of being compact and is used predominantly throughout this book. However, the differential notation $N^{-1}dN/dx$ explicitly presents $f(x)$ for what it is: a density. We thus use it whenever it is important to emphasize the notion of density, most particularly when discussing differential cross sections and correlation functions in Chapters 8 and 9, and toward the definition of histograms introduced in §4.6.

2.6 Functions of Random Variables

Consider a measurement of the momentum, p, of particles produced in proton–proton collisions at some fixed energy. Given the production of particles is a stochastic (random) phenomenon, the momentum of the measured particles can be regarded as a random quantity. Assuming all produced particles are pions, one can determine their energy based on the relativistic relation, $E = \sqrt{p^2c^2 + m^2c^4}$, where m is the mass of the pion and c the speed of

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

light. Since p is a random variable, E is consequently also a random variable. This is true in general: functions of random variables are themselves random variables and it is often of interest to determine or characterize the probability density of such random variables.

Let us first introduce a continuous function, $q(x)$, of a single continuous random variable X. Since X is random, the application of $q(x)$ on observed values x yields random values, and it is thus legitimate to consider these as instances of a random variable Q. Our goal shall be to determine the PDF of Q given a PDF for X.

Let $f(x)$ be the PDF of X. By definition, it represents the probability of observing X in the range $[x, x + dx]$.[5] An observation is an event, and the specific variable used to represent this event should thus be immaterial. This means the probability of observing that event is **conserved** or **invariant** when the variable representing the event is changed or transformed. Let us then consider the same event from the point of view of the variables X and Q. Within the sample space of X, the probability of the event may be expressed

$$f(x)\,dx = \text{Probability an event takes place in } [x, x + dx], \tag{2.56}$$

whereas from within the sample space of Q, one has

$$g(q)\,dq = \text{Probability an event takes place in } [q(x), q(x + dx)], \tag{2.57}$$

where we introduced $g(q)$ as the PDF of the random variable Q. These two expressions represent the same subset of events and must then be equal. In order to find a relation between $g(q)$ and $f(x)$, let us write the function $q(x + dx)$ as a truncated Taylor series:

$$q(x + dx) = q(x) + \left.\frac{dq}{dx}\right|_x dx + O(2). \tag{2.58}$$

In the limit $dx \to 0$, the quantity dq may then be written

$$dq = q(x + dx) - q(x) = \left.\frac{dq}{dx}\right|_x dx. \tag{2.59}$$

Now, given the expressions (2.56) and (2.57) are equal, one can write

$$g(q) = f(x(q)) \left|\frac{dq}{dx}\right|^{-1}. \tag{2.60}$$

Since the function $g(q)$ represents a probability density, one ensures it is positive definite by using the absolute value $|dq/dx|$ in Eq. (2.60). Additionally, note that if $x(q)$ is multivalued, one must include all values of x that map onto a specific value of q.

As a simple example of a multivalued problem, consider the function $q(x) = x^2$ with inverse $x = \pm\sqrt{q}$, and

$$\left|\frac{dx}{dq}\right| = \frac{1}{2\sqrt{q}}.$$

[5] In this context, the interpretation of densities in terms of limiting frequency or degree of belief is somewhat immaterial and we carry out the discussion in terms of the frequentist interpretation for convenience, but the reasonings and results are identical within the Bayesian interpretation.

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

Given there are two roots with equal contributions, one must include a factor of 2 in Eq. (2.60). One thus gets

$$g(q) = \frac{f(x(q))}{\sqrt{q}}. \tag{2.61}$$

The concept of probability density is readily extended to functions of multiple random variables in §2.8, but first it is useful and convenient to introduce commonly used properties of PDFs.

2.7 PDF Characterization

While a PDF carries the maximum amount of information on the behavior of a random variable and the phenomenon or physical quantity it represents, it is often desirable, convenient, and at times sufficient to reduce or transform this information into a set of appropriately chosen properties or functions. For a known PDF, these properties are uniquely defined and can usually be calculated exactly. If the PDF is unknown or partially known, the properties must be estimated based on functions of sampled (measured) data known as **statistics** and **estimators**.[6] Several types of properties are of interest toward the characterization of PDFs and for the modeling of data. We first introduce and discuss, in this section, the notions of **α-point**, **mode**, **expectation value**, **moments**, **centered moments**, and **standardized moments**. The notions of **characteristic function** and **moment-generating functions** are introduced in §2.10 whereas **cumulants** are defined in §2.13. The notions of **covariance** and **factorial moments** are discussed in §2.9 and §10.2 respectively, after the introduction of multivariate random functions in §2.8.

2.7.1 α-Point and Median

The concept of **alpha-point**, also called **quantile of order** α, and noted α-point, is introduced by defining a quantity x_α that demarcates the probability of x being smaller than α:

$$F(x_\alpha) = \alpha \text{ with } 0 \leq \alpha \leq 1. \tag{2.62}$$

The corresponding value x_α is thus the inverse:

$$x_\alpha \equiv F^{-1}(\alpha). \tag{2.63}$$

The value $x_{1/2}$, called the **median**, is a special case of the α-point commonly used to estimate the typical value of a random variable, given there is a 50% probability that measured values of x are smaller. The uniform and Gaussian PDFs shown in Figure 2.2 are symmetric about $x = 0$. Their $1/2$-point is consequently $x_{1/2} = 0$ and the probabilities of x being smaller or larger than 0 are equal.

[6] A formal definition of the notion of a statistics is presented in §4.3.

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

For a discrete random variable x_i, with probability $p(x_i)$, the cumulative distribution

$$F(x) = \sum_{x_i \leq x} p(x_i) \tag{2.64}$$

spans discrete values only. The α-point determined by Eq. (2.63) is thus strictly exact only for discrete values x_α.

2.7.2 Mode

The **mode** of a PDF $f(x)$ is defined as the value of the random variable x for which the PDF has a maximum.

A PDF with a single maximum is called a **unimodal distribution**. The position of the mode may be obtained either by inspection or by finding the extremum of the distribution, in other words, by finding the value x where the PDF has a null first derivative with respect to x. Given that

$$f'(x) \equiv \frac{df(x)}{dx}, \tag{2.65}$$

the mode of the distribution is thus

$$x_{\text{mode}} = f'^{-1}(0). \tag{2.66}$$

PDFs with two or more maxima are known as **bimodal** and **multimodal**, respectively. As for unimodal distributions, their modes can be obtained either by direct inspection or by finding the zeros of the function $f'(x)$.

Figure 2.3 illustrates the concepts of **median**, **mode**, and **mean** of a unimodal PDF as well as the notion of a bimodal PDF.

2.7.3 Expectation Value (Mean)

The expectation value, noted E[x], of a continuous random variable distributed according to a PDF $f(x)$ is defined as

$$\mathrm{E}[x] \equiv \int_{-\infty}^{\infty} x f(x)\, dx \equiv \mu. \tag{2.67}$$

This expression is commonly called **population mean**, **mean**, or **average value** of x. In this book, depending on the context, it will also be denoted as μ, μ_x, or $\langle x \rangle$, where the brackets $\langle\ \rangle$ refer to a population or domain average:

$$\mathrm{E}[x] \equiv \mu \equiv \mu_x \equiv \langle x \rangle. \tag{2.68}$$

The expression E[x] is also commonly referred to as first moment, noted μ_1', of the PDF $f(x)$ for reasons that will become obvious in the following discussion. Brackets [x] rather than parentheses (x) are used to indicate that E[x] is not a function of x but rather represents a certain value of x determined by the shape of the distribution $f(x)$. Indeed, it is a property of the PDF itself and thus not a function of a particular value of x. If $f(x)$ is a strongly peaked function, then the mean is likely to be near the mode of the function. However,

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

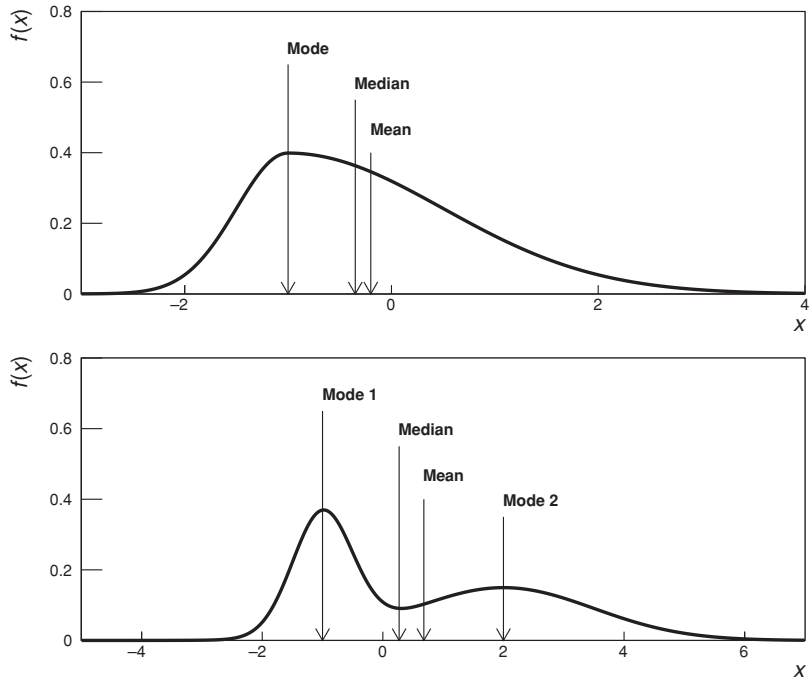

Fig. 2.3 (Top) Illustration of the concepts of median, mode, and mean for a single mode distribution. (Bottom) Example of a bimodal distribution.

for bimodal or multimodal functions, the mean is typically not representative of a specific peak, unless perhaps one peak strongly dominates the others.

As an example of calculation of the expectation value of a function, let us evaluate the mean of the uniform distribution $p_u(x)$ defined in the range $a \leq x \leq b$ according to

$$p_u(x|a,b) = \begin{cases} \frac{1}{b-a} & \text{for } a \leqslant x \leqslant b, \\ 0 & \text{elsewhere,} \end{cases} \tag{2.69}$$

where the denominator $b - a$ ensures the proper normalization of the distribution when $p_u(x|a, b)$ is integrated over $\mathbf{R}$. The mean of x is defined according to

$$\langle x \rangle \equiv \mu_x \equiv \int_{-\infty}^{\infty} p_u(x|a,b) x \, dx. \tag{2.70}$$

Substituting the definition, Eq. (2.69), for $p_u(x|a, b)$, and proceeding with the integration, one finds

$$\langle x \rangle = \frac{1}{b-a} \int_a^b x \, dx, \tag{2.71}$$

$$= \frac{1}{b-a} \left. \frac{x^2}{2} \right|_a^b, \tag{2.72}$$

$$= \frac{a+b}{2}, \tag{2.73}$$

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

which corresponds to the middle of the interval $[a, b]$, and is thus indeed representative of typical values of x determined by $p_u(x|a, b)$.

The notion of expectation value is quite general. In fact, instead of the expectation value of x, one can calculate the expectation value of any functions of x. Let us here denote such a function as $q(x)$. Clearly, since x is a random variable, so shall be $q(x)$. As in §2.6, let $g(q)$ denote the PDF of q. By definition, the expectation value of q shall be

$$\mathrm{E}[q] \equiv \int_{-\infty}^{\infty} q g(q)\, dq. \tag{2.74}$$

Recall from §2.6 that the probability of x being in the interval from x to $x + dx$ must be equal to the probability of q being in the interval from $q(x)$ to $q(x + dx)$. The preceding expectation value may then be written

$$\mathrm{E}[q] = \int_{-\infty}^{\infty} q(x) f(x) \frac{dx}{dq}\, dq = \int_{-\infty}^{\infty} q(x) f(x)\, dx. \tag{2.75}$$

The expectation value of the function $q(x)$, given the PDF $f(x)$, is thus equal to the inner product of $q(x)$ by $f(x)$.

2.7.4 Moments, Centered Moments, and Standardized Moments

A special and important case of the function $q(x)$ involves powers of x. One defines the nth **algebraic moment** (also simply called the nth moment) of the PDF $f(x)$, denoted μ'_n, as

$$\mu'_n \equiv \mathrm{E}[x^n] = \int_{-\infty}^{\infty} x^n f(x)\, dx. \tag{2.76}$$

Obviously, the mean, μ, is a special case of Eq. (2.76) and corresponds to the first ($n = 1$) moment μ'_1.

It is also convenient to consider moments relative to the mean μ. These are defined according to

$$\mu_n \equiv \mathrm{E}[(x - \mathrm{E}[x])^n] = \int_{-\infty}^{\infty} (x - \mu)^n f(x)\, dx, \tag{2.77}$$

and are named **centered moments** or **moments about the mean**. Hereafter, we use the expression **centered moment** exclusively.

A first special and important case to consider is the second centered moment μ_2. It corresponds to the variance, noted $\mathrm{Var}[x]$, of the PDF,

$$\mu_2 \equiv \mathrm{Var}[x] \equiv \mathrm{E}[(x - \mathrm{E}[x])^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\, dx = \sigma^2, \tag{2.78}$$

where σ corresponds to what is commonly known as the **standard deviation** of the PDF. Note that the notation for the variance varies across texts: many authors use the notations $V[x]$, $V(x)$, or $\mathrm{Var}(x)$. In this text, we use the [ ] notation to emphasize that the variance is a functional of the PDF, that is, a property of the PDF rather than, strictly speaking, a function of x. Additionally, depending on the context of our discussions, we shall use

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

several alternative notations for the variance of x as follows:

$$\mu_2 \equiv \text{Var}[x] \equiv \sigma^2 \equiv \sigma_x^2 \equiv \langle (x-\mu)^2 \rangle \equiv \langle \Delta x^2 \rangle, \tag{2.79}$$

where $\Delta x \equiv x - \mu$, and the brackets $\langle\ \rangle$ here again refer to a population or domain average.

One can then show (see Problem 2.4) that the variance may be expressed in terms of the second and first moments:

$$\mu_2 \equiv \text{E}[x^2] - \mu^2 = \mu'_2 - \mu^2. \tag{2.80}$$

One also readily verifies that the variance Var[x] measures the spread of x about its mean value μ, as illustrated in Figure 2.4 with four selected Gaussian distributions. The Gaussian distribution $p_G(x|\mu, \sigma)$, formally introduced in §3.9, is defined according to

$$p_{\text{G}}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \tag{2.81}$$

where the factor $\sqrt{2\pi}\sigma$ enables proper normalization of the distribution when it is integrated over **R**. Distributions shown in Figure 2.4 are symmetric about the origin and thus have a mean $\langle x \rangle$ equal to zero. Setting $\mu = 0$ in Eq. (2.81), one proceeds to calculate the variance of x according to

$$\text{Var}[x] = \text{E}[x^2] - \mu^2 = \text{E}[x^2], \tag{2.82}$$

$$= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2\sigma^2}\right) x^2\, dx. \tag{2.83}$$

Substituting $z = x/\sigma$, the preceding expression becomes

$$\text{Var}[x] = \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{z^2}{2}\right) z^2\, dz. \tag{2.84}$$

The integral is readily determined to equal $\sqrt{2\pi}$ from basic definite integral tables. The variance of the Gaussian distribution is then

$$\text{Var}[x] = \sigma^2. \tag{2.85}$$

Comparing the distributions plotted in Figure 2.4 for selected values of σ, one finds, indeed, that a Gaussian with a large spread in x features a large variance, whereas a narrow Gaussian has a small variance. This result can be readily extended to distributions of arbitrary shapes: broadly distributed PDFs have a large variance whereas narrowly distributions have a small variance.

It is often useful to compare the higher moments, $n > 2$, of a distribution to the standard deviation. This may be accomplished with **standardized moments**, μ_k^{std}, defined as ratios of the kth centered moments and the kth power of the standard deviation:

$$\mu_k^{\text{std}} = \frac{\mu_k}{\sigma^k}. \tag{2.86}$$

The standard moments hence correspond to kth moments normalized with respect to the standard deviation. The power of k is required in the normalization given moments scale as x^k. This implies the standardized moments are scale invariant; in other words, rescaling

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

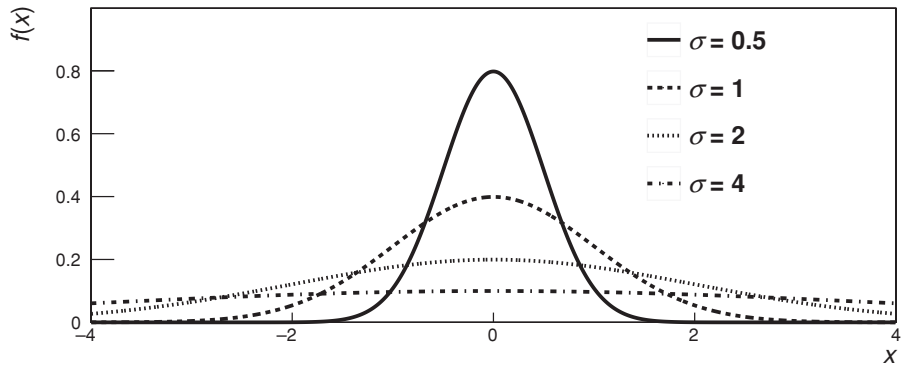

Fig. 2.4 Four Gaussian PDFs (see §3.9) with mean, $\mu = 0$, and standard deviations σ of 0.5, 1.0, 2.0, and 4.0. The variance measures the breadth of a distribution. The distribution shown as a solid line is the narrowest and has the smallest variance. The other distributions are wider and thus have larger variances.

of the variable x by an arbitrary factor α leads to changes in the kth moment and the standard deviation by a factor α^k, but their ratio is invariant, that is, independent of α. The standardized moments μ_k^{std} are dimensionless numbers for the same reason. Also note that the first standardized moment vanishes because the first moment about the mean is null, while the second standardized moment equals unity because the second moment about the mean is the variance. The third and fourth standardized moments are called skewness and kurtosis, respectively. These are discussed in the next two sections.

2.7.5 Skewness

The **skewness** of a distribution is commonly denoted γ_1 or Skew[x] in the literature. It is formally defined as the third standardized moment of a distribution:

$$\gamma_1 \equiv \text{Skew}[x] \equiv \mu_3^{\text{std}} = \frac{\mu_3}{\sigma^3}. \tag{2.87}$$

Skewness is essentially a measure of the asymmetry of a distribution. Consider, for instance, the distributions shown in Figure 2.5. The solid line curve displays a Gaussian distribution (defined in §3.9) with mean, $\mu = 0$, and width, $\sigma = 0.6$. It is by construction symmetric about its mean and therefore has null skewness. The dash and dotted curves are constructed as two juxtaposed half Gaussian distributions. Their peaks are both at $x = 0$ (same as for the black curve) but the left and right widths of the dash (dotted) curve are 0.6 (2.0) and 1.5 (0.6), respectively. Focusing on the dash curve, one finds the right side of the distribution tapers differently than the left side. These tapering sides, called low and high side tails, provide a visual means for determining the sign of the skewness of a distribution. The skewness is typically negative if the low side (left) tail is longer than the high side (right) tail. It is then said to be left-skewed (dotted curve in Figure 2.5). If the high side (right) tail is longer, the distribution is said to be right-skewed (dashed curve). In a skewed (unbalanced, lopsided) distribution, the mean is farther out in the long tail than is the median. Distributions with zero skewness, such as normal distributions, are symmetric

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

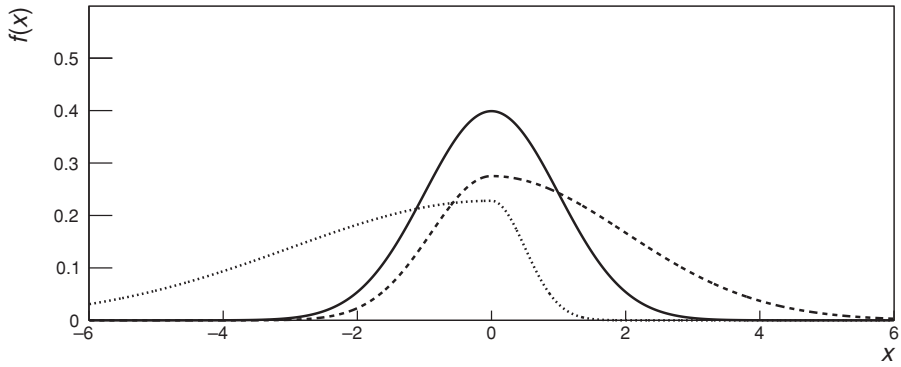

Fig. 2.5 Illustration of the notion of skewness. The solid line curve is by construction symmetric and has, as such, zero skewness. The dashed and dotted curves have longer high side and low side tails and consequently have positive and negative skewness, respectively.

about their mean: in effect, their mean equals their median and mode. Note, however, that it is not sufficient for the mean of a distribution to be right (left) of the median to conclude it is right (left) skew. Multimodal and discrete distributions, in particular, may violate this simple expectation.

Karl Pearson[7] (1857–1936) suggested several alternative measures of skewness:

$$\text{Pearson mode skewness: } \frac{\mu - \text{mode}}{\sigma}, \tag{2.88}$$

$$\text{Pearson's 1st skewness coefficient: } 3\frac{\mu - \text{mode}}{\sigma}, \tag{2.89}$$

$$\text{Pearson's 2nd skewness coefficient: } 3\frac{\mu - \text{median}}{\sigma}. \tag{2.90}$$

These are, however, not frequently used to characterize data in nuclear and particle physics.

2.7.6 Kurtosis

Two definitions of **kurtosis** are commonly used in modern statistical literature. The first corresponds to the old kurtosis and is often noted Kurt[x]. It is defined as the fourth standardized moment of a distribution. As such, it corresponds to the ratio of the fourth centered moment by the fourth power of the standard deviation:

$$\text{Kurt}_{\text{old}}[\text{x}] \equiv \frac{\mu_4}{\sigma^4}. \tag{2.91}$$

Modern kurtosis is usually called **excess kurtosis**. Denoted γ_2, it is defined as the ratio of the fourth cumulant by the square of the second cumulant (see §2.13.1 for the definition

[7] Influential English mathematician generally credited for establishing the discipline of mathematical statistics.

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

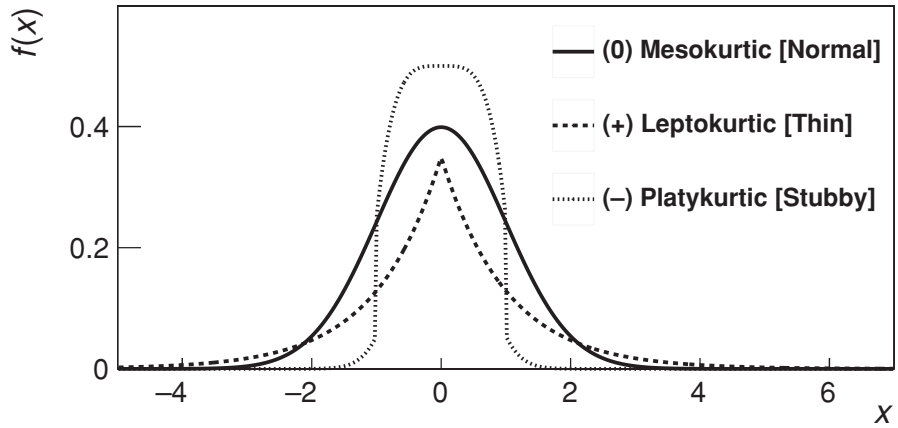

Fig. 2.6 Illustration of the notion of kurtosis. The solid line is a Gaussian distribution and has zero excess kurtosis. The dashed curve has a peaked distribution with long tails and thus has a positive kurtosis while the dotted curved has a flat top with short tails and hence is characterized by a negative kurtosis.

of cumulants):

$$\gamma_2 = \frac{\kappa_4}{\kappa_2^2} = \frac{\mu_4}{\sigma^4} - 3. \tag{2.92}$$

The "minus 3" conveniently makes kurtosis of the Gaussian distribution equal to zero. More importantly, the definition in terms of cumulants implies the (excess) kurtosis of a sum of n independent random variables x_i is equal to the sum of the kurtosis of these n variables divided by n^2. This can be written (see Problem 2.6):

$$\gamma_2 \left[\sum_{i=1}^{n} x_i \right] = \frac{1}{n^2} \sum_{i=1}^{n} \gamma_2 [x_i]. \tag{2.93}$$

Such a simple scaling by n^2 does not arise with the "old" kurtosis definition.

Figure 2.6 schematically illustrates the notion of kurtosis. A high kurtosis distribution has a sharper peak and longer, fatter tails than a Gaussian distribution, while a low kurtosis distribution has a more rounded peak and shorter, thinner tails. Distributions with zero excess kurtosis are called **mesokurtic**, or **mesokurtotic**. The most obvious example of a mesokurtic distribution is the Gaussian distribution (see §3.9). A few other well-known distributions can be mesokurtic, depending on their parameter values. For example, the binomial distribution (see §3.1) is mesokurtic for $p = 1/2 \pm \sqrt{1/12}$. A distribution with positive excess kurtosis is called **leptokurtic**, or **leptokurtotic**. A leptokurtic distribution has a more acute peak around its mean and fatter tails than a Gaussian distribution (narrower peak and more probable extreme values). Examples of leptokurtic distributions include the Laplace distribution and the logistic distribution; such distributions are sometimes called super-Gaussian. A distribution with negative excess kurtosis is called **platykurtic**, or **platykurtotic**. The shape of a platykurtic distribution features a lower, wider peak around the mean (i.e., a lower probability than a normally distributed variable of values

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

near the mean) and thinner tails; in other words, extreme values (both smaller and larger than the mean) have a larger probability than a normally distributed variable. The uniform distribution (see §3.4) is a prime example of a platykurtic distribution. Another example involves the Bernoulli distribution (see §3.1), with $p = 1/2$, obtained, for example, for the number of times one obtains "heads" when flipping a coin (i.e., a coin toss), for which kurtosis is -2. Such distributions are sometimes termed sub-Gaussian.

The notion of kurtosis is not as frequently used as the notion of variance but nonetheless remains of interest in general to characterize the shape of a distribution. It finds specific applications in nuclear physics with the study of net charge fluctuations and the determination of the charge susceptibility of the quark gluon plasma (see §11.3.3).

2.7.7 Credible Range

Perhaps the most basic way to characterize a set of data is to describe the **range** it covers. The range of the data, as the word suggests, is simply the difference between the highest and lowest observed values of a random variable. It is consequently straightforward to determine, although it may be somewhat misleading for distributions with long low or high side tails (with low probability). In such cases, the range is subject to large fluctuations when dealing with small samples and is thus rather unreliable in characterizing the bulk of a distribution. It may then be preferable to use the notion of **interquartile range** instead, which is evaluated as the difference between the higher and lower quartiles. The lower and higher quartiles correspond to 1/4-point and 3/4-point, respectively.

The notions of **deciles** and **percentiles** are also commonly used to report placements within a distribution as fractions of a population, sample, or distribution with values smaller or equal to a given decile or percentile. For instance, students receiving a 99 percentile score on a physics exam have good reasons to be proud because they were among the top 1% of all test-takers.

The dispersion or spread of a distribution may also be reported by quoting its full width at half maximum (or FWHM), as illustrated in Figure 2.7. The FWHM presents the advantage, relative to the standard deviation, of being fairly immune to the effects associated with low or high side tails. As such, it is useful to characterize the width of the main body of unimodal distributions. It is easy to verify (see Problem 2.9) that the FWHM of a Gaussian distribution is

$$\text{FWHM} = 2.35\sigma \qquad \text{Gaussian distribution.} \tag{2.94}$$

2.8 Multivariate Systems

Practical scientific problems are rarely limited to measurements of a single (random) variable. Especially in particle and nuclear physics, modern experiments involve measurements of large numbers of physical variables simultaneously. At the detector level, measured quantities amount to voltages produced by sensors, which can eventually be interpreted

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

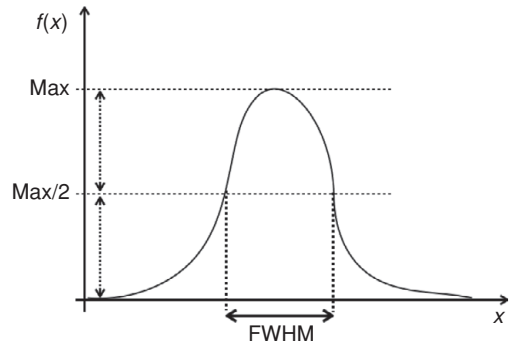

Fig. 2.7 Illustration of the notion of full width at half maximum (FWHM).

as particle positions, momenta, or energies, and possibly a host of other physical quantities. Nuclear collisions produce varying number of particles with random values of momentum, energy, or even particle species. The number of physical variables involved in nucleus–nucleus interactions may vary from a handful in soft proton–proton interactions to thousands in head-on Pb on Pb collisions at the Large Hadron Collider. A fraction of these variables may be correlated, while others may be completely independent. However, it is usually not known a priori which variables are statistically independent and which others are correlated. One is then compelled, at least conceptually, to formulate the notion of multivariate (i.e., multiple variables) probability densities that encompass all measured variables. However, for simplicity's sake and without loss of generality, we will first examine the notion of multivariate probability densities using two variables only.

This discussion can be carried out equivalently in terms of the frequentist and Bayes approaches to probability. Here, we will adopt the Bayesian perspective and use the language of probability as logic. See [67] for an introduction of the same concepts based on sets and the frequentist approach.

2.8.1 Joint Probability Density Functions

Let us consider a system involving two continuous random observables (or model parameters). Let X represent the hypothesis (or statement) that the first observable lies in the range $[x, x + dx]$ and Y that the second lies in the range $[y, y + dy]$ given some prior information about the system, I. The conjunction (AND) of the two hypotheses, X, Y, is true if both hypotheses are true. The quantity $p(X, Y|I)$ then expresses the degree of belief that the hypotheses X and Y might be true jointly (i.e., simultaneously). Given X and Y are continuous hypotheses, the conjunction X, Y is also a continuous hypothesis, and $p(X, Y|I)$ thus corresponds to a **joint probability density function**

$$p(X, Y|I) = \lim_{\Delta x, \Delta y \to 0} \frac{p(x \leq X < x + \Delta x, y \leq Y < y + \Delta y|I)}{\Delta x \Delta y}, \tag{2.95}$$

that expresses the degree of belief the variables X and Y be found jointly in the intervals $[x, x + dx]$ and $[y, y + dy]$, respectively.

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

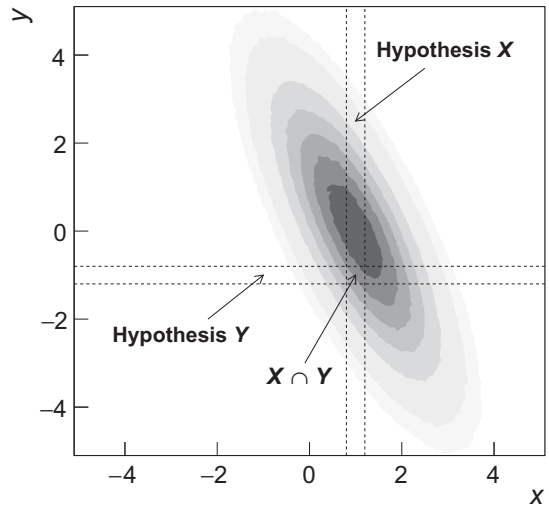

Fig. 2.8 Contour plot of the joint probability $f(x, y)$ of two variables X and Y generated according to a 2D Gaussian model. The horizontal dashed lines delimit hypothesis Y while the vertical dashed lines represent the range of hypothesis X. The probability of a pair (x, y) to lie within the square given by the conjunction of X and Y is equal to $f(x, y)\Delta x\Delta y$, with $\Delta x\Delta y$ being the surface area of the square.

Evidently, one could also consider alternative hypotheses that the observables X and Y are found in other intervals. Summing probabilities of hypotheses spanning the entire hypothesis space, **H**, yields unity. One thus gets the normalization condition

$$\iint_{\mathbf{H}} p(X, Y|I)\, dX\, dY = 1. \tag{2.96}$$

The notion of joint probability density is illustrated in Figure 2.8 which shows a probability density function of two variables, x and y, in the form of an iso-contour plot (the boundaries between different shades of gray delineate loci of equal probability density). While x and y are both random variables with Gaussian distributions, their values are not independent of one another: large values of x tend to be accompanied by small (i.e., negative) values of y and small (i.e., negative) values of x tend to be observed in conjunction with large values of y. The two variables are then said to be **correlated**. A measure of the degree of correlation between two variables can be obtained with the notion of covariance discussed in §2.9.3.

2.8.2 Marginal Probability Density Functions

Let us now assume that the PDF $p(X, Y|I)$ is given. It is obviously also of interest to determine the probability density $p(X|I)$ corresponding to the probability that the hypothesis X is true irrespective of other hypotheses. We next show $p(X|I)$ is readily obtained by integration of $p(X, Y|I)$ over all hypotheses Y. This operation is referred to as **marginalization** in the statistics literature.

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

To demonstrate this result, let us first assume the hypotheses Y are discrete and may be labeled Y_i, with $i = 1, \ldots, n$. Let us further assume the Y_i are collectively exhaustive, $\sum_{i=1}^{n} Y_i = 1$, and mutually exclusive, $Y_i, Y_j = 0$ for $i \neq j$. One can then write

$$p\left(\sum_{i=1}^{n} Y_i | I\right) = 1. \tag{2.97}$$

Let us then consider the probability $p(X, \sum_{i=1}^{n} Y_i | I)$ asserting the degree of belief that X and $\sum_{i=1}^{n} Y_i$ are jointly true. Application of the product yields

$$p\left(X, \sum_{i=1}^{n} Y_i | I\right) = p\left(\sum_{i=1}^{n} Y_i | I\right) p\left(X | \sum_{i=1}^{n} Y_i, I\right) \tag{2.98}$$

$$= 1 \times p\left(X | \sum_{i=1}^{n} Y_i, I\right). \tag{2.99}$$

But since $\sum_{i=1}^{n} Y_i = 1$, the conjunction of this and the prior I is simply I, and one gets

$$p\left(X, \sum_{i=1}^{n} Y_i | I\right) = p(X|I), \tag{2.100}$$

which is the result we are looking for. We must now express the left-hand side of this expression in terms of probabilities $p(X, Y_i|I)$. This is readily accomplished by noting that a conjunction (AND) can be distributed onto a disjunction (OR), that is,

$$X, \sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} X, Y_i. \tag{2.101}$$

Since the Y_i are mutually exclusive, the propositions X, Y_i are also mutually exclusive, and one can use the extended sum rule, Eq. (2.40), to obtain

$$p(X|I) = p\left(X, \sum_{i=1}^{n} Y_i | I\right) = \sum_{i=1}^{n} p(X, Y_i|I). \tag{2.102}$$

Our derivation was based on discrete statements Y_i. Let us now assume there is an infinite number of such statements (collectively exhaustive and mutually exclusive); we must then replace the sum by an integral and the probabilities by densities, and one gets the sought for result:

$$p(X|I)\, dx = \left(\int_{\mathbf{H}} p(X, Y|I)\, dy\right) dx \tag{2.103}$$

or simply

$$p(X|I) = \int_{\mathbf{H}} p(X, Y|I)\, dy. \tag{2.104}$$

The probability density function $p(X|I)$ is said to be the marginal probability density of $p(X, Y|I)$, or alternatively, one can say that $p(X, Y|I)$ has been marginalized or that the "uninteresting" parameter Y has been eliminated by **marginalization**. Quite obviously,

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

given conjunctions commute, that is, $A, B = B, A$, one can achieve the marginalization of X in the same fashion:

$$p(Y|I) = \int_{\mathbf{H}} p(X, Y|I)\, dx. \tag{2.105}$$

The use of probability notation $p(X|I)$, $p(Y|I)$, $p(X, Y|I)$, and so forth, may become rapidly tedious. It is thus convenient to introduce an alternative notation based on more traditional function notation (common in the frequentist interpretation). For instance, representing the density $p(X, Y|I)$ by a function $f(x, y)$, it is common to denote marginal probabilities $p(X|I)$ as $f_x(x)$ and one writes

$$f_x(x) = \int f(x, y),\, dy, \tag{2.106}$$

$$f_y(y) = \int f(x, y),\, dx, \tag{2.107}$$

where the integrals are taken over the domains of the integrated variables.

Experimentally, the PDF $f(x, y)$ may be estimated using a two-dimensional histogram involving a very large number of measurements of pairs (x, y). The marginal PDFs $f_x(x)$ and $f_y(y)$ may then be estimated from projections of the two-dimensional histogram onto axes x and y, respectively, as illustrated in Figure 2.9. Two-dimensional and multidimensional histograms and their projections are formally discussed in §4.6.

2.8.3 Conditional Probability Density Functions

Given a known joint PDF $p(X, Y|I)$ and the marginal PDFs $p(X|I)$ and $p(Y|I)$ (or in more traditional notation: $f(x, y)$, $f_x(x)$, and $f_y(y)$), it is also of interest to evaluate the probability density, $p(X|Y, I)$, corresponding to the probability density that X is true when Y is known to be true. Since X and Y are continuous hypotheses, this amounts to the conditional probability density for X to be in the interval $[x, x + dx]$ given that Y is known to be in $[y, y + yx]$. Applying the product rule onto $p(X, Y|I)$, we readily get

$$p(X, Y|I) = p(Y|I)p(X|Y, I). \tag{2.108}$$

Rearranging, and substituting the expression obtained in Eq. (2.105) for $p(Y|I)$, one gets

$$p(X|Y, I) = \frac{p(X, Y|I)}{p(Y|I)} \tag{2.109}$$

$$= \frac{p(X, Y|I)}{\int_{\mathbf{H}} p(X, Y|I)\, dx}, \tag{2.110}$$

which is the **Conditional Probability Density Function** of X given Y. Evidently, the commutativity of the conjunction operation implies one can also write

$$p(Y|X, I) = \frac{p(X, Y|I)}{p(X|I)} \tag{2.111}$$

$$= \frac{p(X, Y|I)}{\int_{\mathbf{H}} p(X, Y|I)\, dy}, \tag{2.112}$$

which is the **Conditional Probability Density Function** of Y given X.

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

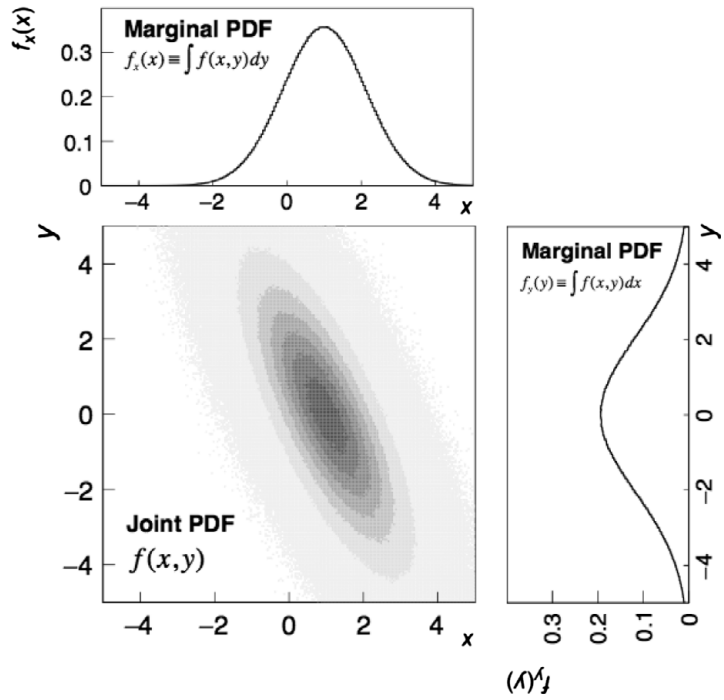

Fig. 2.9 Illustration of the notion of marginal probability. The contour plot presents the joint probability $f(x, y)$ of two variables x and y defined according to a 2D Gaussian model. The right and top panels show the marginal probabilities $f_y(y)$ and $f_x(x)$ obtained by integrating $f(x, y)$ over x and y, respectively.

Here again, it also convenient to introduce a somewhat more traditional function notation, and one writes

$$h_x(x|y) \equiv \frac{f(x,y)}{f_y(y)} = \frac{f(x,y)}{\int f(x',y)\,dx'} \tag{2.113}$$

$$h_y(y|x) \equiv \frac{f(x,y)}{f_x(x)} = \frac{f(x,y)}{\int f(x,y'),\,dy'}, \tag{2.114}$$

which defines $h_x(x|y)$ as a conditional probability density of x given y and $h_y(y|x)$ as a conditional probability density of y given x.

We stress that the conditional PDF $p(X|Y, I)$, or equivalently $h_x(x|y)$, must be regarded as a function of a single variable x in which y is treated as a constant value. It expresses the probability (density) of getting a certain value of x given a specific value of y has already been observed, and conversely for $p(Y|X, I)$.

We discuss in §4.6.3 how to obtain estimates of $h_y(y|x)$, experimentally, from a two-dimensional histogram by projection onto the x-axis of a slice taken at a specific value of y. In general, different choices of the constant y, for instance y_1 and y_2, lead to different conditional probabilities noted $h_x(x|y_1) \neq h_x(x|y_2)$, as illustrated in Figure 2.10. Note, however, that given both functions are PDFs, they both satisfy the normalization

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

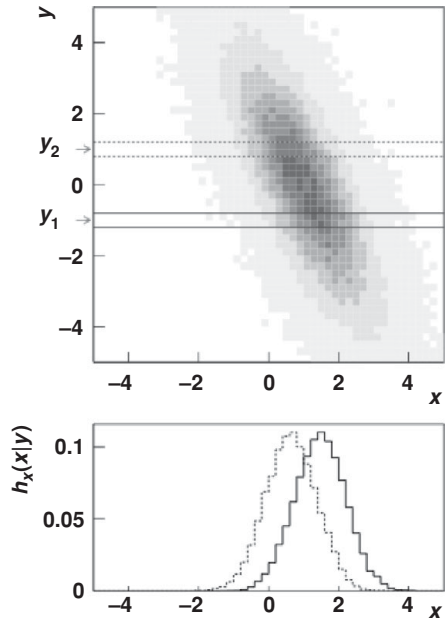

Fig. 2.10 Illustration of the notion of Conditional Probability Density Function. The scatterplot presents the joint probability density $f(x, y)$ of two variables x and y randomly generated according to a 2D Gaussian model. The bottom panel show the conditional probabilities $h_x(x|y_1)$ (solid line) and $h_x(x|y_2)$ (dashed line) obtained by integrating $f(x, y)$ along y in the ranges $[y_1 - \epsilon, y_1 + \epsilon]$ and $[y_2 - \epsilon, y_2 + \epsilon]$, respectively (with $\epsilon = 0.2$).

condition:

$$\int_{-\infty}^{-\infty} h_x(x|y_1)\, dx = 1, \tag{2.115}$$

$$\int_{-\infty}^{-\infty} h_x(x|y_2)\, dx = 1. \tag{2.116}$$

2.8.4 Bayes' Theorem and Probability Densities

Combining Eqs. 2.109 and 2.111 (or equivalently Eqs. 2.113 and 2.114), one arrives at an expression of Bayes' theorem in terms of the marginal and conditional PDFs of continuous variables x and y:

$$p(X|Y, I) = \frac{p(Y|X, I)p(X|I)}{p(Y|I)} \tag{2.117}$$

where the functions $p(X|Y, I)$, $p(Y|X, I)$, $p(X|I)$, and $p(Y|I)$ are probability density functions. Using the alternative function notation introduced in the preceding text, this may be written:

$$h_x(x|y) = \frac{h_y(y|x) f_x(x)}{f_y(y)}. \tag{2.118}$$

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

Continuing with this notation, we note that Eqs. (2.113, 2.114) can also be written

$$f(x,y) = h_y(y|x) f_x(x) = h_x(x|y) f_y(y), \tag{2.119}$$

One then obtains the expressions

$$f_x(x) = \int_{-\infty}^{-\infty} h_x(x|y) f_y(y)\, dy, \tag{2.120}$$

$$f_y(y) = \int_{-\infty}^{-\infty} h_y(y|x) f_x(x)\, dx, \tag{2.121}$$

which correspond to the law of total probability applied towards the determination of marginal probabilities $f_x(x)$ and $f_y(y)$.

Indeed, given $f_y(y)$ and $h_x(x|y)$, one can use Eq. (2.120) to derive the density $f_x(x)$. As we shall discuss in §12.3, Eq. (2.120) can be used, in particular, to fold and unfold smearing and efficiency effects associated with instrumental artifacts provided a model of the detector performance is available. One can also use Eq. (2.120), or Eq. (2.121), to account for physical effects. For instance, the function $f_y(y)$ might represent the momentum spectrum of a full particle jet (composed of neutral and charged particles) produced in elementary nuclear interactions, and $h_x(x|y)$ could model the probability of measuring charged jet momenta x given a full jet momentum y. The function $f_x(x)$, calculated with Eq. (2.120), would then represent the momentum distribution of charged jets.

Next, recall from Eq. (2.11) that if two hypotheses A and B are independent, the probability of their conjunction must satisfy $p(A, B|I) = p(A|I)p(B|I)$. Two continuous variables x and y can thus be considered **statistically independent** if their joint PDF factorizes as follows:

$$f(x,y) = f_x(x) f_y(y) \quad \text{(statistical independence)}. \tag{2.122}$$

This, in turn, implies that the conditional PDFs $h_y(y|x)$ and $h_x(x|y)$ are the same for all values of x and y. Indeed, substituting the preceding expression for $f(x,y)$ in Eqs. (2.113) and (2.114), one gets

$$h_y(y|x) = \frac{f_x(x) f_y(y)}{f_x(x)} = f_y(y) \quad \text{(statistical independence)}, \tag{2.123}$$

$$h_x(x|y) = \frac{f_x(x) f_y(y)}{f_y(y)} = f_x(x) \quad \text{(statistical independence)}, \tag{2.124}$$

from which we conclude that if two variables x and y are statistically independent, their conditional probability densities equal their marginal densities. This implies that having knowledge of one variable does not influence the probability of the other in any way. Conversely, finding that conditional densities $h_y(y|x)$ and $h_x(x|y)$ depend on x and y, respectively, would be a sure indication that the two variables are not statistically independent.

2.8.5 Extension to $m > 2$ random variables

The preceding discussion can be readily extended to measurements involving any number m of random variables, x_i, $i = 1, \ldots, m$. The probability of measuring the m variables in

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

ranges $[x_i, x_i + dx_i]$ defines the joint PDF $f(x_1, x_2, \ldots, x_m)$. Given this PDF is a function of multiple variables, one can define several marginal probabilities $f_{x_i}(x_i)$:

$$f_{x_i}(x_i) = \int \cdots \int f(x_1, x_2, \ldots, x_m)\, dx_1 dx_2 \ldots dx_{i-1} dx_{i+1} \ldots dx_m. \tag{2.125}$$

One can also define marginal PDFs that are functions of several variables. For two variables, e.g., x_1 and x_2, one gets

$$f_{x_1,x_2}(x_1, x_2) = \int \cdots \int f(x_1, x_2, \ldots, x_m)\, dx_3 dx_4 \ldots dx_m, \tag{2.126}$$

which can easily be generalized to any two (or more) variables.

The extension of conditional probabilities to multiple variables proceeds similarly. For instance, the conditional probability density of getting x_1 given values x_2,..., x_m may be written

$$h_{x_1}(x_1|x_2, \ldots x_m) = \frac{f(x_1, x_2, \ldots, x_m)}{\int f(x_1', x_2, \ldots, x_m)\, dx_1'}. \tag{2.127}$$

This expression can be generalized to obtain the conditional PDF of finding several variables. For instance, the conditional PDF of x_1, x_2, given x_3,..., x_m is given by

$$h_{x_1,x_2}(x_1, x_2|x_3, \ldots x_m) = \frac{f(x_1, x_2, \ldots, x_m)}{\int f(x_1', x_2', \ldots, x_m)\, dx_1' dx_2'}. \tag{2.128}$$

The methods based on Bayes' theorem and the law of total probability presented earlier for functions of two variables can be readily extended to calculate marginal and conditional PDFs of several variables (see Problem 2.11).

2.8.6 Multivariate Functions of Random Variables

Equipped with the notion of multivariate probability densities introduced earlier in §2.8, we proceed to discuss multivariate functions of random variables.

Let $q(x_1, \ldots, x_n)$ represent a function of multiple random variables $x_1, x_2, \ldots, x_n$.[8] Obviously, given the x_i are continuous random variables, the value $q(x_1, \ldots, x_n)$ may also be regarded as a random variable characterized by a probability density $g(q)$. The probability (density) of specific values q is determined in part by the function itself and in part by the likelihood of getting combinations of $x_1, x_2, \ldots, x_n$ that yield that given value. In turn, this likelihood is determined by the joint probability density $f(x_1, x_2, \ldots, x_n)$ of the variables. The probability of observing a value q in the range $[q, q + dq]$ may then be obtained by summing all relevant combinations of values of $x_1, \ldots, x_n$, that is, all such values that yield a value q in that range. This may be written

$$g(q)\, dq = \int_{d\Omega} f(x_1, x_2, \ldots, x_n)\, dx_1 dx_2 \cdots dx_n, \tag{2.129}$$

[8] Again here, the variables x_i may represent continuous hypotheses (Bayesian interpretation) or the outcome of some series of measurements (frequentist interpretation).

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

where the volume element $d\Omega$ encloses the region in $x_1, x_2, \ldots, x_n$ space between the two hypersurfaces $q(x_1, x_2, \ldots, x_n) = q$ and $q(x_1, x_2, \ldots, x_n) = q + dq$. The size and shape of $d\Omega$ are obviously determined by the function $q(x_1, \ldots, x_n)$ itself. The preceding integral is thus generally nontrivial and its evaluation may require numerical methods, including Monte Carlo methods discussed in §13.2. However, there are several interesting cases that can be handled analytically, a few of which we examine in the following because they are frequently encountered in data analysis problems.

First, consider a case where two independent random variables X and Y are distributed according to PDFs $f_x(x)$ and $f_y(y)$, respectively. Suppose we wish to calculate the PDF $f_z(z)$ corresponding to a function $z(x, y)$ of the two variables. Because X and Y are independent variables, the joint PDF $f(x, y)$ is simply the product of the functions $f_x(x)$ and $f_y(y)$. The determination of $f_z(z)$ thus reduces to the relatively simple integral

$$f_z(z)\,dz = \int_{d\Omega} f_x(x) f_y(y)\,dx dy, \tag{2.130}$$

where the domain of integration $d\Omega$ includes all combinations of x and y satisfying $z \equiv z(x, y)$.

Let us proceed with three specific examples, starting with the integration for $z = x \pm y$. One writes

$$f_z(z)\,dz = \int_{-\infty}^{\infty} f_x(x)\,dx \int_{z\pm x}^{(z+dz)\pm x} f_y(y)\,dy. \tag{2.131}$$

The inner integral is carried over an infinitesimal range dz across which the function f_y does not change. One gets

$$f_z(z)\,dz = dz \int_{-\infty}^{\infty} f_x(x) f_y(z \pm x)\,dx, \tag{2.132}$$

which implies the density $f_z(z)$ may be written

$$f_z(z) = \int_{-\infty}^{\infty} f_x(x) f_y(z \pm x)\,dx. \tag{2.133}$$

Alternatively, reversing the order of integrations, one finds

$$f_z(z) = \int_{-\infty}^{\infty} f_x(z \pm y) f_y(y)\,dy. \tag{2.134}$$

This result can also be obtained using δ-functions. We can enforce the requirement $z = x \pm y$ by inserting a δ-function $\delta\,(z - (x \pm y))$ into Eq. (2.130) while carrying out the integration over all possible values of both x and y. The integration over one of the variables then becomes trivial and one gets

$$f_z(z) = \int_{-\infty}^{\infty} f_x(x)\,dx \int_{-\infty}^{\infty} f_y(y)\delta(z - (x \pm y))\,dy, \tag{2.135}$$

$$= \int_{-\infty}^{\infty} f_x(x) f_y(z \pm x)\,dx. \tag{2.136}$$

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

This expression is commonly written $f_z = f_x \otimes f_y$ and is called the **Fourier convolution** of f_x and f_y. Note that in practical situations, complications may arise because measurements of the function f_x and f_y may be limited to ranges $x_{\min} \leq x \leq x_{\max}$ and $y_{\min} \leq y \leq y_{\max}$, beyond which the functions do not necessarily vanish but cannot be measured. The evaluation of the integral must consequently be limited to the boundaries of the measurement exclusively. Such cases are encountered, for instance, in measurements of correlation functions discussed in Chapter 10. Fourier convolutions are also commonly encountered in smearing or resolution modeling of the response of detectors (see, e.g., §12.2.3).

A generalization of the preceding result for a function $z(x_1, x_2, \ldots, x_n)$ which is a linear combination of the variables x_i

$$z = \sum_{i=1}^{n} c_i x_i, \tag{2.137}$$

yields

$$f_z(z) = \int_{-\infty}^{\infty} dx_1 \cdots \int_{-\infty}^{\infty} dx_{n-1} f\left(x_1, \ldots, x_{n-1}, \left[z - \tfrac{1}{c_n} \textstyle\sum_{i=1}^{n-1} c_i x_i\right]\right). \tag{2.138}$$

Next, consider the integration of Eq. (2.130) for $z = xy$, which we carry out by inserting the δ-function $\delta(z - xy)$ into the convolution integral:

$$f_z(z) = \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy f_x(x) f_y(y) \delta(z - xy). \tag{2.139}$$

The δ-function $\delta(g(x))$ may be written

$$\delta(g(x)) = \sum_i \frac{\delta(x - x_i)}{|g'(x_i)|}, \tag{2.140}$$

where $g'(x_i)$ is the derivative of $g(x)$ with respect to x evaluated at the roots x_i of $g(x)$. In the case under consideration, we have $g(x) = z - xy$. There is a single root $x_o = z/y$ and $|g'(x_o)| = |y|$. One then gets

$$f_z(z) = \int_{-\infty}^{\infty} f_x(z/y) \frac{f_y(y)}{|y|} dy. \tag{2.141}$$

This function f_z is known as the **Mellin convolution** of f_x and f_y. Mellin convolutions are very useful in physics. They provide, in particular, a convenient technique toward the calculation of jet fragmentation functions and DGLAP evolution (see, e.g., [150] § 20).

As a third case, we consider the function $z = x/y$. The convolution integral may then be written

$$f_z(z) = \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy f_x(x) f_y(y) \delta(z - x/y), \tag{2.142}$$

which yields after integration over x:

$$f_z(z) = \int_{-\infty}^{\infty} |y| f_x(yz) f_y(y)\, dy. \tag{2.143}$$

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

2.9 Moments of Multivariate PDFs

We proceed to extend the notion of expectation value introduced in §2.7.3 to include functions of several variables. For instance, the mean or expectation values of a function $q(x_1, x_2, \ldots, x_n)$ may be written

$$\mu_q \equiv \mathrm{E}[q(\vec{x})] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} q(\vec{x}) f(\vec{x})\, dx_1 \cdots dx_n, \tag{2.144}$$

where for the sake of simplicity, we have introduced a vector $\vec{x} = (x_1, \ldots, x_n)$ to denote the dependence over the variables $x_1, x_2, \ldots, x_n$ and the function $f(\vec{x})$ corresponds to the joint PDF $f(x_1, x_2, \ldots, x_n)$ of the variables $x_1, x_2, \ldots, x_n$. Let us examine the calculation of the expectation value, Eq. (2.144), for selected and particularly relevant cases of the function $q(\vec{x})$.

2.9.1 First-Order Moments of x_i

In general, the function $q(\vec{x})$ may consist of linear or nonlinear functions of the variables x_i. However, as a first and simple case, it is useful to choose any of the n random variables x_i and calculate their first moments μ_i according to

$$\mu_i \equiv \mathrm{E}[x_i] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} x_i f(\vec{x})\, dx_1 \cdots dx_n. \tag{2.145}$$

In this context, the symbols μ_i correspond to the means of each of the random variables x_i and should not be confused with the higher moments of a single variable introduced earlier in this chapter. It is convenient to represent the means μ_i as a vector of n elements

$$\vec{\mu} = (\mu_1, \mu_2, \ldots, \mu_n), \tag{2.146}$$

which one can interpret as the mean of the PDF in the full n-dimension space spanned by the random variables x_i.

2.9.2 Variance of x_i

Next, consider the variance of the function $q(\vec{x})$. By definition, one has

$$\begin{aligned} \sigma_q^2 \equiv \mathrm{Var}[q] &= \mathrm{E}[\big(q(\vec{x}) - \mu_q\big)^2] \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \big(q(\vec{x}) - \mu_q\big)^2 f(\vec{x})\, dx_1 \cdots dx_n. \end{aligned} \tag{2.147}$$

Choosing once again $q(\vec{x}) = x_i$, one obtains the variance of the multivariate PDF relative to each of the n random variables x_i:

$$\begin{aligned} \sigma_i^2 \equiv \mathrm{Var}[x_i] &= \mathrm{E}[(x_i - \mu_i)^2] \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (x_i - \mu_i)^2 f(\vec{x})\, dx_1 \cdots dx_n. \end{aligned} \tag{2.148}$$

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

2.9.3 Covariance of Two Variables x_i and x_j

It is also useful to consider the expectation value of products such as $x_i x_j$ for $i, j = 1, \ldots, n$, and $i \neq j$. We are more specifically interested in centered moments of two variables x_i and x_j relative to their respective means and define the **covariance**, $\mathrm{Cov}[x_i, x_j]$, of variables x_i and x_j, with $i \neq j$ as:

$$\mathrm{Cov}[x_i, x_j] \equiv \mathrm{E}[(x_i - \mu_i)\,(x_j - \mu_j)], \tag{2.149}$$

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (x_i - \mu_i)\,(x_j - \mu_j)\, f(\vec{x})\, dx_1 \cdots dx_n. \tag{2.150}$$

The covariance of two random variables measures the degree to which the variables are **correlated**, or **covarying**. Consider, for instance, the covariance of two variables x_1 and x_2 for a density $f(x_1, x_2)$. The preceding expression becomes

$$\mathrm{Cov}[x_1, x_2] = \int (x_1 x_2 - x_1 \mu_2 - \mu_1 x_2 + \mu_1 \mu_2)\, f(x_1, x_2)\, dx_1 dx_2, \tag{2.151}$$

where μ_1 and μ_2 are the mean values of variables x_1 and x_2, respectively. Splitting the four terms of the integrand, we get

$$\begin{aligned} \mathrm{Cov}[x_1, x_2] = &\int x_1 x_2 f(x_1, x_2)\, dx_1 dx_2 - \mu_2 \int x_1 f(x_1, x_2)\, dx_1 dx_2 \\ &- \mu_1 \int x_2 f(x_1, x_2)\, dx_1 dx_2 + \mu_1 \mu_2 \int f(x_1, x_2)\, dx_1 dx_2, \end{aligned} \tag{2.152}$$

and noting that the integrals of the second and third terms are simply μ_1 and μ_2, respectively, while the integral of the last term is unity by virtue of the normalization of $f(x_1, x_2)$, Eq. (2.152) reduces to

$$\mathrm{Cov}[x_1, x_2] = \int x_1 x_2 f(x_1, x_2)\, dx_1 dx_2 - \mu_1 \mu_2 \tag{2.153}$$

A null covariance indicates the variables x_1 and x_2 may be statistically independent. Indeed, $\mathrm{Cov}[x_1, x_2] = 0$ means one can write

$$\int x_1 x_2 f(x_1, x_2)\, dx_1 dx_2 = \mu_1 \mu_2 = \int x_1 f_{x_1}(x_1)\, dx \int x_2 f_{x_2}(x_2)\, dx_2. \tag{2.154}$$

which in turn suggests $f(x_1, x_2) = f_{x_1}(x_1) f_{x_2}(x_2)$ as expected, if the variables x_1 and x_2 are statistically independent. However, we will see later in this section that the factorization (and statistical independence) is not strictly guaranteed by a null covariance, $\mathrm{Cov}[x_1, x_2] = 0$.

The interpretation of the notion of covariance is best illustrated with the practical examples of joint probability densities of two random variables X and Y presented in Figure 2.11. The joint distributions shown in panels (a–c), are defined as product of two

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

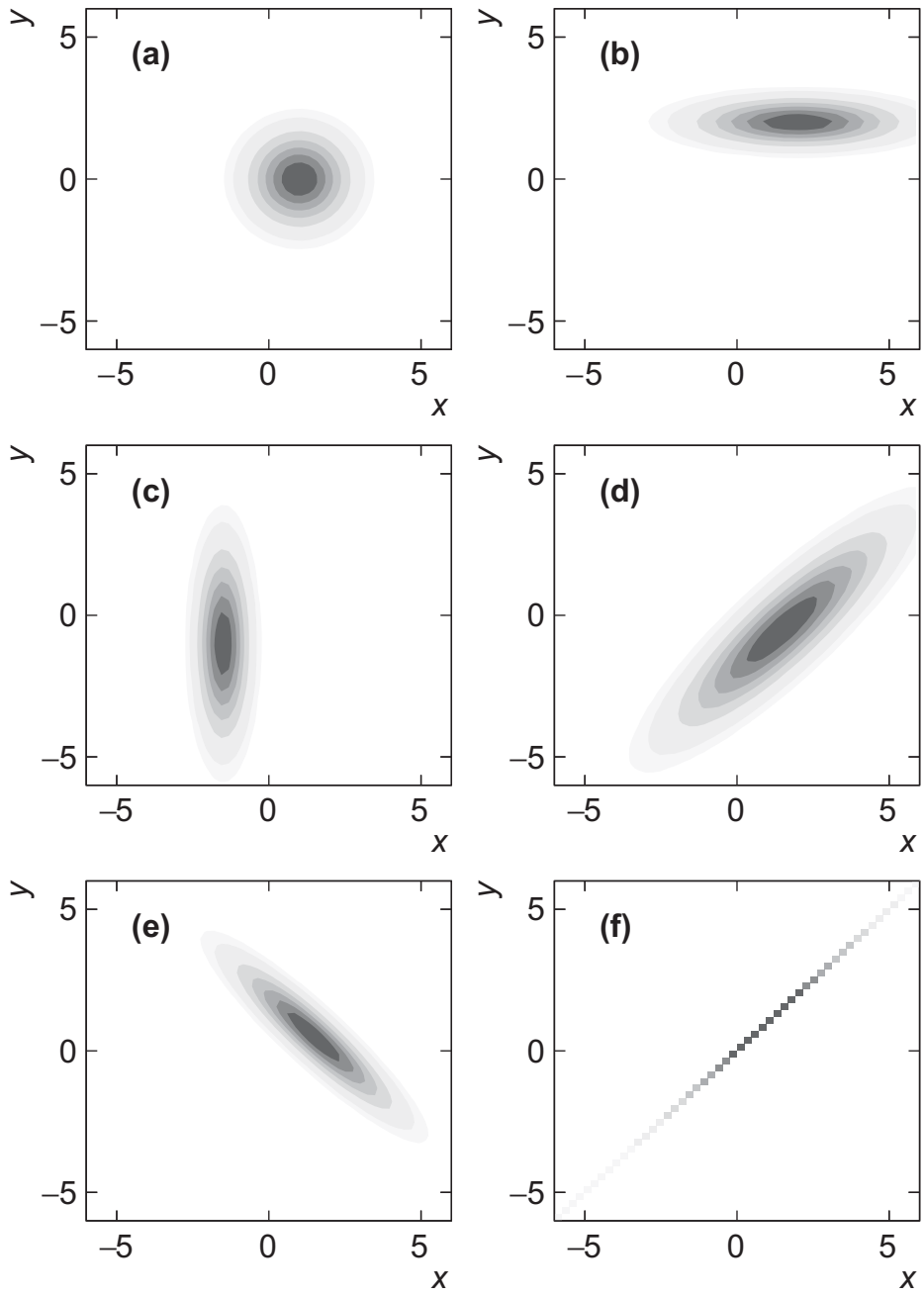

Fig. 2.11 Illustration of the notion of covariance of two random continuous variables. Panels (a), (b), and (c) present examples of uncorrelated variables whereas panels (d) and (e) show examples of fluctuations with positive and negative covariance, respectively. Panel (f) presents a special case where the covariance is maximal (Pearson coefficient is unity). See text for details.

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

independent Gaussians according to

$$p_{a-c}(x,y) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(x-\mu_x)^2}{2\sigma_x^2}\right] \times \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(y-\mu_y)^2}{2\sigma_y^2}\right] \tag{2.155}$$

with means μ_x, μ_y and standard deviations σ_x, σ_y, for variables x and y, respectively. The joint PDF shown in (a) features equal standard deviations $\sigma_x = \sigma_y$ whereas those shown in (b) and (c) are defined with $\sigma_x > \sigma_y$ and $\sigma_x < \sigma_y$, respectively. Given their definition as a product of independent Gaussian distributions, one readily verifies that the covariance of x and y for these distributions may be written

$$\begin{aligned}\mathrm{Cov}[x,y] = &\frac{1}{\sqrt{2\pi}} \int \exp\left[-\frac{(x-\mu_x)^2}{2\sigma_x^2}\right](x-\mu_x)\, dx \\ &\times \frac{1}{\sqrt{2\pi}} \int \exp\left[-\frac{(y-\mu_y)^2}{2\sigma_y^2}\right](y-\mu_y)\, dy\end{aligned}$$

and is thus null because the expectation values of x and y equal μ_x and μ_u, respectively, by definition of the Gaussian distribution.

The distributions shown in panels (d–f) represent correlated joint distributions of the variables X and Y defined according to

$$\begin{aligned} x &= \mu_x + r_1 + \alpha r_2, \\ y &= \mu_y + r_1 - \alpha r_2, \end{aligned} \tag{2.156}$$

where r_1 and r_2 represent independent Gaussian distributed random variables

$$\begin{aligned} p_1(r_1) &= \tfrac{1}{\sqrt{2\pi}} \exp\left[-\tfrac{r_1^2}{2\sigma_1^2}\right], \\ p_2(r_2) &= \tfrac{1}{\sqrt{2\pi}} \exp\left[-\tfrac{r_2^2}{2\sigma_2^2}\right], \end{aligned} \tag{2.157}$$

with null expectation values, $\mathrm{E}[r_1] = \mathrm{E}[r_1] = 0$, and standard deviations σ_1 and σ_2, respectively. The parameter α is set to unity in panels (d-e) and to zero in panel (f). Panel (d) illustrates a case with $\sigma_1 \gg \sigma_2$ implying fluctuations of r_1 are much larger than those of r_2 and thus dominate the values of both x and y, whereas panel (d) features a case with $\sigma_1 \ll \sigma_2$ in which fluctuations of r_2 are much larger than those of r_1. Dominant fluctuations of r_1 (over those of r_2) lead to covarying values of x and y: when x increases, so does y on average as seen in panel (d). In contrast, dominant fluctuations of r_2 lead to anti-correlated values of x and y: as x increases, y tends to decrease (on average).

In order to calculate the covariance of x and y for these three distributions, we first express r_1 and r_2 in terms of x and y by inversion of Eq. (2.156) with $\alpha = 1$:

$$\begin{aligned} r_1 &= \tfrac{1}{2}\left(x + y - \mu_x - \mu_y\right), \\ r_2 &= \tfrac{1}{2}\left(x - y - \mu_x + \mu_y\right). \end{aligned} \tag{2.158}$$

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

On can then calculate the joint PDFs of x and y in terms of the joint PDF of r_1 and r_2 as follows:

$$p_{d-f}(x,y) = \frac{d^2N}{dxdy} = \frac{d^2N}{dr_1dr_2}\left|\frac{\partial\,(r_1,r_2)}{\partial\,(x,y)}\right|. \tag{2.159}$$

The Jacobian $|\partial\,(r_1,r_2)\,/\partial\,(x,y)\,| = 1/2$ is readily calculated from Eq. (2.158). One then obtains

$$p_{d-f}(x,y) \equiv \frac{1}{2}\frac{1}{\sqrt{2\pi}\sigma_1}\exp\left(-\frac{r_1^2}{2\sigma_1^2}\right)\frac{1}{\sqrt{2\pi}\sigma_2}\exp\left(-\frac{r_2^2}{2\sigma_2^2}\right). \tag{2.160}$$

Substituting values for r_1 and r_2 from Eq. (2.158), one gets

$$\begin{aligned} p_{d-f}(x,y) = \frac{1}{2}\frac{1}{\sqrt{2\pi}\sigma_1}\exp\left[-\frac{\left(x+y-\mu_x-\mu_y\right)^2}{8\sigma_1^2}\right] \\ \times\frac{1}{\sqrt{2\pi}\sigma_2}\exp\left[-\frac{\left(x-y-\mu_x+\mu_y\right)^2}{8\sigma_2^2}\right], \end{aligned} \tag{2.161}$$

which cannot be factorized in terms of functions of x and y independently, owing to the fact that the two variables are correlated. Calculation of the covariance of x and y for these distributions is best accomplished in terms of the variables r_1 and r_2 as follows:

$$\mathrm{Cov}[x,y] = \int p_{d-f}\,(x-\mu_x)\left(y-\mu_y\right)\,dxdy, \tag{2.162}$$

$$= \int \frac{d^2N}{dr_1dr_2}\,(r_1+r_2)\,(r_1-r_2)\,dr_1dr_2, \tag{2.163}$$

$$= \int p_1(r_1)p_2(r_2)\left(r_1^2-r_2^2\right)\,dr_1dr_2, \tag{2.164}$$

$$= \int p_1(r_1)r_1^2dr_1 - \int p_1(r_2)r_2^2dr_2, \tag{2.165}$$

$$= \sigma_1^2-\sigma_2^2, \tag{2.166}$$

where in the third line we used the fact that d^2N/dr_1dr_2 factorizes, in the fourth line the normalization to unity of the Gaussian distribution, and in the last line, we substituted the variance σ_1^2 and σ_2^2 of r_1 and r_2, respectively. One finds that for $\sigma_1 > \sigma_2$, when the fluctuations of r_1 are larger than those of r_2, that the covariance of x and y is positive, reflecting that an increase of x is on average accompanied by an increase of y, while for $\sigma_1 < \sigma_2$, the covariance is negative, corresponding to the reverse behavior, that is, an increase of x is on average accompanied by a decrease of y. Lastly, we remark that in the case of panel (f), the variables x and y are perfectly correlated ($\alpha = 0$), thereby implying that the covariance of x and y equals the variance of x.

2.9.4 Covariance Matrix V_{ij}

For a multivariate distributions, $f(\vec{x})$, with $\vec{x} = (x_1, x_2, \ldots, x_n)$, it is useful to extend the notion of covariance to also encompass the notion of variance. This enables the definition

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

of a covariance matrix, V_{ij}, as follows:

$$V_{ij} \equiv \mathrm{E}[(x_i - \mu_i)\,(x_j - \mu_j)], \tag{2.167}$$

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (x_i - \mu_i)\,(x_j - \mu_j)\, f(\vec{x})\, dx_1 \cdots dx_n, \tag{2.168}$$

where all values $i, j = 1, \ldots, n$ are allowed, including $i = j$.

More generally, given two functions $q_1(\vec{x})$ and $q_2(\vec{x})$ of n variables $\vec{x} = (x_1, x_2, \ldots, x_n)$, the covariance $\mathrm{Cov}[q_1, q_2]$ is calculated as follows:

$$\begin{aligned} \mathrm{Cov}[q_1, q_2] &\equiv \mathrm{E}[(q_1 - \mu_{q_1})\,(q_2 - \mu_{q_2})], \\ &= \mathrm{E}[q_1 q_2] - \mu_{q_1}\mu_{q_2}, \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} q_1 q_2 g(q_1, q_2)\, dq_1 dq_2 - \mu_{q_1}\mu_{q_2}, \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} q_1(\vec{x}) q_2(\vec{x}) f(\vec{x})\, dx_1 \cdots dx_n - \mu_{q_1}\mu_{q_2}, \end{aligned} \tag{2.169}$$

where $g(q_1, q_2)$ is the joint probability density for q_1 and q_2 while $f(\vec{x})$ is the joint PDF for $\vec{x} = (x_1, x_2, \ldots, x_n)$. Obviously, $\mathrm{Cov}[q_1, q_2]$ is by construction invariant under permutation of the variables q_1 and q_2:

$$\mathrm{Cov}[q_1, q_2] = \mathrm{Cov}[q_2, q_1]. \tag{2.170}$$

This implies the matrix V_{ij} defined earlier is *symmetric* by construction and its elements satisfy

$$V_{ij} = V_{ji}. \tag{2.171}$$

2.9.5 Correlation Coefficients ρ_{ij}

The off-diagonal matrix elements V_{ij}, $i \neq j$, measure the degree of correlation between the random variables x_i and x_j. Given each of the variables in general exhibits a finite variance, $V_{ii} > 0$, it is useful to quantify the covariances of two variables x_i and x_j relative to their respective variances by introducing correlation coefficients ρ_{ij} defined as

$$\rho_{ij} = \frac{V_{ij}}{\sqrt{V_{ii}V_{jj}}} = \frac{V_{ij}}{\sigma_i \sigma_j}. \tag{2.172}$$

These coefficients are commonly referred to as Pearson correlation coefficients in the literature. It can be shown they are bound in the range $-1 \leq \rho_{ij} \leq 1$ (see Problem 2.12).

2.9.6 Interpretation and Special Cases

The covariance $\mathrm{Cov}[x, y]$ characterizes the degree of correlation between the random variables X and Y. There are phenomena such that when x is greater than its mean μ_x, y is also likely (on average) to be larger than its mean μ_y, and conversely, when $x < \mu_x$ so does $y < \mu_y$. The differences $x - \mu_x$ and $y - \mu_y$ are together positive or negative, which

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

implies the two quantities are positively correlated, $\mathrm{Cov}[x, y] > 0$. Other phenomena show a negative correlation, $\mathrm{Cov}[x, y] < 0$. In this case, when $x < \mu_x$, one is more likely to observe $y > \mu_y$, and conversely when $x > \mu_x$, one observes $y < \mu_y$. In still other phenomena, there is no preference $y < \mu_y$ or $y > \mu_y$ when x is smaller or larger than its mean; the two variables are thus uncorrelated. This occurs when the joint PDF $f(x, y)$ can be factorized, $f(x, y) = f(x)f(y)$. Then clearly one gets $\mathrm{Cov}[x, y] = 0$:

$$f(x, y) = f(x)f(y) \quad \text{implies } \mathrm{E}[x, y] = \mathrm{E}[x]\mathrm{E}[y]. \tag{2.173}$$

However, the converse is not necessarily true. A null covariance does not always guarantee the joint PDF factorizes. Indeed, there can be cases where the integral, Eq. (2.149), is null even though the joint PDF does not factorize. For instance, consider the case $y = x^2$ for a uniform PDF, $f(x)$, defined in the range $[-1, 1]$. Obviously, y and x are then perfectly correlated by construction. Yet one finds

$$\begin{aligned} \mathrm{Cov}\,[y, x] &= \mathrm{Cov}\left[x^2, x\right], \\ &= E\left[x^3\right] - E\left[x^2\right] E\,[x], \\ &= 0 - 0 \times E\left[x^2\right], \\ &= 0. \end{aligned} \tag{2.174}$$

This implies that although a null covariance, $\mathrm{Cov}\,[y, x] = 0$, in general suggests the variables x and y are statistically independent, one must be cautious to reach this conclusion too hastily and verify against pathological cases such as the foregoing one.

2.10 Characteristic and Moment-Generating Functions

The characteristic function (CF), noted $\phi_x(t)$, of a real valued random variable is defined, in the complex plane, as the inverse Fourier transform of the probability density function (PDF) of this variable while the Moment-Generating Function (MGF), denoted $M_x(t)$, is defined on the set of real numbers as the Laplace transform of the PDF. By construction, a CF (or MGF) is uniquely defined by its PDF, and conversely, a PDF is uniquely defined by a CF (or MGF). CF and MGF may then be used as alternative definitions and characterizations of the behavior of a random variable. CF and MGF are particularly useful toward the calculation of the moments of PDFs. They also are useful in the calculation of other PDF properties, such as their limiting behavior, and in the demonstration of several important theorems of probability and statistics. The CF and MGF of a PDF have very similar definitions based on Fourier and Laplace transforms and can often be used interchangeably. It is important to note, however, that the MGF of a PDF does not always exist. This is particularly the case when one or more of the expectation values $\mathrm{E}[x^k]$ of a PDF diverge (e.g., the Breit–Wigner PDF). By contrast, one can show that the characteristic function of a real-valued PDF always exists, although it does not entail the existence of its moments.

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

2.10.1 Definitions of $\phi_x(t)$ and $M_x(t)$

The characteristic functions $\phi_x(t)$ of a random variable x with PDF $f(x)$ is defined as the expectation value of the function e^{itx}

$$\phi_x(t) = E\left[e^{itx}\right] = \int_{-\infty}^{\infty} e^{itx} f(x)\, dx, \tag{2.175}$$

and as such essentially corresponds to the inverse Fourier transform of the function $f(x)$. The MGF, $M_x(t)$, is defined as the Laplace transform of $f(x)$ for values t limited to the set of real numbers:

$$M_x(t) = E\left[e^{tx}\right] = \int_{-\infty}^{\infty} e^{tx} f(x)\, dx, \qquad t \in \mathrm{R} \tag{2.176}$$

There appears to be very little difference between the two definitions. We will, however, see that the expectation value $E\left[e^{ts}\right]$ of some PDFs diverges and $M_x(t)$ consequently does not always exist. At the same time, one can show that $\phi_x(t)$ always exists although it may be of limited use in practice.

Equation (2.175) establishes a one-to-one relationship between the PDF $f(x)$ of a variable x and its characteristic function $\phi_x(t)$. This implies that if $f(x)$ is not known a priori, it may be determined on the basis of the Fourier transform of its characteristic function $\phi(x)$:

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \phi_x(t)\, dt. \tag{2.177}$$

This property is useful, as we shall see in Section 2.10.3, to determine the PDF of sums of variables and to derive several important results in probability and statistics.

2.10.2 Calculation of the Moments μ'_k of a PDF

Let us first show that both $\phi_x(t)$ and $M_x(t)$ can be used to calculate the moments of a PDF.

Expansion of the exponential e^{tx} in series yields

$$e^{tx} = 1 + tx + \frac{t^2x^2}{2!} + \frac{t^3x^3}{3!} + \cdots + \frac{t^nx^n}{n!} + \cdots \tag{2.178}$$

The expectation value $E\left[e^{tx}\right]$ may then be written

$$E\left[e^{tx}\right] = 1 + tE\left[x\right] + \frac{t^2E\left[x^2\right]}{2!} + \frac{t^3E\left[x^3\right]}{3!} + \cdots + \frac{t^nE\left[x^n\right]}{n!} + \cdots \tag{2.179}$$

given t is here considered a "parameter" of the function. Since the expectation values $E\left[x^k\right]$ correspond to the moments μ'_k of the PDF, one may then write

$$E\left[e^{tx}\right] = 1 + t\mu' + \frac{t^2\mu'_2}{2!} + \frac{t^3\mu'_3}{3!} + \cdots + \frac{t^n\mu'_n}{n!} + \cdots \tag{2.180}$$

Derivatives of this expression with respect to t evaluated at $t = 0$ yield the moments μ'_k of the PDF:

$$\mu'_k = \left.\frac{d^k}{dt^k} E\left[e^{tx}\right]\right|_{t=0} = \left.\frac{d^k}{dt^k} M_x(t)\right|_{t=0}. \tag{2.181}$$

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

The same reasoning for $\phi_x(t)$ yields (see Problem 2.18)

$$\mu'_k = i^{-k} \frac{d^k}{dt^k} \phi_x(t) \bigg|_{t=0}. \tag{2.182}$$

We demonstrate the use of this expression for the calculation of the moments of the Gaussian distribution. The characteristic function of the Gaussian distribution is obtained by calculating the integral

$$\phi(t) = E\left[e^{itx}\right] = \int_{-\infty}^{\infty} \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} e^{itx}\, dx. \tag{2.183}$$

Using integration tables, and after some basic algebra, one gets

$$\phi(t) = \exp\left(i\mu t - \tfrac{1}{2}\sigma^2 t^2\right). \tag{2.184}$$

The first moment (i.e., the mean) of the Gaussian PDF is obtained by evaluating the first derivative of Eq. (2.184) at $t = 0$, with respect to t. One gets

$$\mu'_1 = i^{-1} \frac{d}{dt} \phi(t) \bigg|_{t=0} = i^{-1} \left(i\mu - \sigma^2 t\right) \exp\left(i\mu t - \tfrac{1}{2}\sigma^2 t^2\right) \big|_{t=0} = \mu, \tag{2.185}$$

which is the anticipated result. Similarly, calculation of the second moment μ'_2 requires a second derivative of the characteristic function, also evaluated at $t = 0$. One finds

$$\begin{aligned} \mu'_2 &= i^{-2} \frac{d^2}{dt^2} \phi(t) \bigg|_{t=0} \\ &= i^{-2} \left\{ -\sigma^2 \exp\left(i\mu t - \tfrac{1}{2}\sigma^2 t^2\right) + \left(i\mu - \sigma^2 t\right)^2 \exp\left(i\mu t - \tfrac{1}{2}\sigma^2 t^2\right) \right\} \bigg|_{t=0} \\ &= \sigma^2 + \mu^2, \end{aligned} \tag{2.186}$$

which yields

$$\mathrm{Var}[x] = \mu'_2 - (\mu'_1)^2 = \sigma^2, \tag{2.187}$$

also as anticipated. The calculation of higher moments of the Gaussian distribution proceeds in a similar fashion.

The CFs and MGFs of the probability distributions used in this textbook are provided in Tables 3.1 to 3.8, as are moments of these distributions.

Note that although the characteristic function of a PDF always exist, its derivatives, Eq. (2.182), may not. This is, for instance, the case of the Cauchy PDF and, by extension, the Breit–Wigner PDF (see §3.15). One can show that the characteristic function of the Cauchy distribution is $\phi(t) = \exp(-|t|)$. This function is not differentiable at $t = 0$. Its moments consequently do not exist (i.e., they diverge).

2.10.3 Sum of Random Deviates

Imagine one has n independent random variables $x_1,\ldots, x_n$, with respective PDFs $f_1(x_1),\ldots, f_n(x_n)$ and their corresponding characteristic functions $\phi_1(t),\ldots, \phi_n(t)$. Let us construct a

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

new random variable z as the sum of the variables x_1,…, x_n:

$$z = \sum_{i=1}^{n} x_i. \tag{2.188}$$

We wish to determine the probability density function $f(z)$ on the basis of the PDFs $f_1(x_1), \ldots,$ and $f_n(x_n)$. Rather than explicitly integrating the product of the functions $f_i(x_i)$ as in Section 2.8.6, it turns out to be simpler to first calculate the characteristic function $\phi_z(t)$.

$$\phi_z(t) = E\left[e^{itz}\right], \tag{2.189}$$

$$= \int \cdots \int \exp\left(it \sum_{i=1}^{n} x_i\right) f_1(x_1) \cdots f_n(x_n)\, dx_1 \cdots dx_n, \tag{2.190}$$

which may also be written

$$\phi_z(t) = \int e^{itx_i} f_1(x_1)\, dx_1 \cdots \int e^{itx_n} f_n(x_n)\, dx_n, \tag{2.191}$$

$$= \phi_1(t) \cdots \phi_n(t). \tag{2.192}$$

The characteristic function of a random variable z, which is a sum of random variables x_i, is thus simply the product of the characteristic functions of these variables. The PDF $f(z)$ can then be obtained by inverse Fourier transform:

$$f(z) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \phi_z(t) e^{-itz}\, dt, \tag{2.193}$$

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} \phi_1(t) \cdots \phi_n(t) e^{-itz}\, dt. \tag{2.194}$$

As a first example of application of this theorem, let us determine the probability density function, $f(z)$, of a variable z defined as the sum of two Gaussian variables x_1 and x_2 with means μ_1 and μ_2 and widths σ_1 and σ_2, respectively. Using (2.191), we write the characteristic function of the sum of Gaussian variables as

$$\phi_z(t) = \phi_1(t)\phi_2(t), \tag{2.195}$$

where $\phi_1(t)$ and $\phi_2(t)$ are characteristic functions of the Gaussian variables x_1 and x_2, respectively. Substituting the expression of the characteristic function for a Gaussian variable, Eq. (2.184), we find

$$\phi_z(t) = \exp\left[i\mu_1 t - \tfrac{1}{2}\sigma_1^2 t^2\right] \exp\left[i\mu_2 t - \tfrac{1}{2}\sigma_2^2 t^2\right], \tag{2.196}$$

$$= \exp\left[i\,(\mu_1 + \mu_2)\,t - \tfrac{1}{2}\left(\sigma_1^2 + \sigma_2^2\right)t^2\right]. \tag{2.197}$$

Introducing the variables $\mu_z = \mu_1 + \mu_2$ and $\sigma_z = \sqrt{\sigma_1^2 + \sigma_2^2}$, we find that the preceding expression may be rewritten

$$\phi_z(t) = \exp\left[i\mu_z t - \tfrac{1}{2}\sigma_z^2 t^2\right], \tag{2.198}$$

which is itself the expression of the characteristic function of a Gaussian PDF with mean μ_z and σ_z. We have thus demonstrated the notion that the sum of two Gaussian variables yields

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

another Gaussian variable with a mean equal to the sum of the means of the individual variables and a width equal to the sum, in quadrature, of their respective widths.

The preceding calculation can easily be applied to the difference between two Gaussian variables $\Delta x = x_1 - x_2$ (see Problem 2.20). One finds the difference has a mean $\mu_{\Delta x} = \mu_1 - \mu_2$ and a variance $\sigma^2_{\Delta x} = \sqrt{\sigma_1^2 + \sigma_2^2}$. One may also extend the preceding calculation (see Problem 2.21) to a sum of an arbitrary number n of Gaussian deviates x_i with mean μ_i and width σ_i. Note that a similar result applies for a sum of n Poisson variables (see Problem 2.22).

2.10.4 Central Limit Theorem

The **Central Limit Theorem** (CLT) stipulates that the sum X of n independent continuous random variables, x_i, $i = 1, \ldots, n$, taken from distributions of mean μ_i and variance V_i (or σ_i^2), respectively, has a probability density function $f(X)$ with the following properties:

1. The expectation value of X is equal to the sum of the means μ_i,

$$\langle X \rangle \equiv \mathrm{E}[X] = \sum_{i=1}^{n} \mu_i. \tag{2.199}$$

2. The variance of $f(X)$ is given by the sum of the variances σ_i^2,

$$\langle \Delta X^2 \rangle \equiv \mathrm{Var}[X] = \sum_{i=1}^{n} V_i = \sum_{i=1}^{n} \sigma_i^2. \tag{2.200}$$

3. The function $f(X)$ becomes a Gaussian in the limit $n \to \infty$.

The CLT finds applications in essentially all fields of scientific study. Whether discussing the behavior of complex systems, or scrutinizing the details of scientific measurements, one finds that measured variables are in general influenced by a large number of distinct effects and processes. The more complex they are, the larger the number of effects and processes. This means that fluctuations, and consequently measurement errors tend to have a Gaussian distribution. The CLT is thus truly important, and that is why so much emphasis is also given to discussions of Gaussian distributions.

The proof of items 1 and 2 of the CLT is straightforward and is here presented first. The third item is less obvious and discussed next at greater length.

Let us assume the variables x_i follow PDFs $f_i(x_i)$ of mean μ_i and variance V_i. Given X is defined as a sum

$$X = \sum_{i=1}^{n} x_i, \tag{2.201}$$

one can write

$$\mathrm{E}[X] = \mathrm{E}\left[\sum_{i=1}^{n} x_i\right] = \sum_{i=1}^{n} \mathrm{E}[x_i] \tag{2.202}$$

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

since sum operations commute. Thus, insofar as the mean μ_i are defined, one gets item 1 of the theorem

$$\mathrm{E}[X] = \sum_{i=1}^{n} \mu_i. \tag{2.203}$$

The calculation of the variance of $f(X)$ proceeds similarly:

$$\mathrm{Var}[X] = \mathrm{E}\left[(X - \mathrm{E}[X])^2\right], \tag{2.204}$$

$$= \mathrm{E}\left[\left(\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \mu_i\right)^2\right], \tag{2.205}$$

$$= \mathrm{E}\left[\left(\sum_{i=1}^{n} (x_i - \mu_i)\right)^2\right]. \tag{2.206}$$

To calculate the square within the expectation value, note that one can write $\left(\sum_{i=1}^{n} a_i\right)^2 = \sum_{i=1}^{n} a_i^2 + \sum_{i\neq j=1}^{n} a_i a_j$, where diagonal terms are separated from nondiagonal terms. The variance thence becomes

$$\mathrm{Var}[X] = \sum_{i=1}^{n} \mathrm{E}\left[(x_i - \mu_i)^2\right] + \sum_{i\neq j=1}^{n} \mathrm{E}\left[(x_i - \mu_i)(x_j - \mu_j)\right]. \tag{2.207}$$

The expectation value in the first term corresponds to the variance of the variables x_i, whereas the expectation value in the second term yields the covariances of variables x_i and x_j. Since the variables x_i and x_j are assumed to be independent, these covariances vanish, and one is left with the anticipated result

$$\mathrm{Var}[X] = \sum_{i=1}^{n} \mathrm{Var}[x_i]. \tag{2.208}$$

If the measurements are not independent, Eq. (2.199) still holds but Eqs. (2.200, 2.208) do not and Eq. (2.207) must be used instead.

We next turn to the derivation of the third item of the central limit theorem, i.e., the notion that in the large n limit, the sum of n variables follows a Gaussian distribution. It is already clear from the addition theorem presented in §2.10.3 that the CLT applies for n variables x_i that are Gaussian distributed. Our task is now to demonstrate that the theorem holds also for deviates with arbitrary PDFs.

Equation (2.208) tells us that the variance of the sum z is equal to the sum of the variances σ_i^2. Introducing the average variance $\sigma^2 = (\sum_{i=1}^{n} \sigma_i^2)/n$, one thus expects the variance of the sum z to scale as $n\sigma^2$. It is then convenient to introduce

$$y_i = \frac{x_i - \mu_i}{\sqrt{n}}, \tag{2.209}$$

which satisfies $\mathrm{E}[y_i] = 0$ and $\mathrm{Var}[y_i] = \sigma_i^2/n$. We may thus recast the problem of finding the PDF of $z = \sum x_i$ in terms of $z' = \sum y_i$, which by construction shall have a null mean and a variance equal to σ^2. Our task is thus reduced to the calculation of the characteristic

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

function $\phi_{z'}(t)$ given by

$$\phi_{z'}(t) = \prod_{i=1}^{n} \phi_i(t), \tag{2.210}$$

where $\phi_i(t)$ are the characteristic functions of variables y_i. Rather than using specific expressions for the functions $\phi_i(t)$, let us use a generic expansion of the function e^{ity_i} and write

$$\phi_i(t) = \mathrm{E}[e^{ity_i}] = \sum_{m=0}^{\infty} \frac{1}{m!}(it)^m \mathrm{E}[y_i^m]. \tag{2.211}$$

Insertion of this expression in Eq. (2.210) with values from Eq. (2.209) for y_i yields

$$\phi_{z'}(t) = \prod_{i=1}^{n} \left(\sum_{m=0}^{\infty} \frac{i^m t^m}{m! n^{m/2}} \mathrm{E}[(x_i - \mu_i)^m] \right). \tag{2.212}$$

Let us introduce the shorthand notation $V_i^m = \mathrm{E}[(x_i - \mu_i)^m]$ and recall that by construction $V_i^1 = 0$. The calculation of the foregoing product of sums, although tedious, is simple if terms are grouped in powers of t. One gets at the lowest order in the following expression:

$$\phi_{z'}(t) = 1 - \frac{t^2}{2n} \sum_{j=1}^{n} V_j^2 - \frac{it^3}{3! n^{3/2}} \sum_{j=1}^{n} V_j^3 \tag{2.213}$$

$$+ \frac{t^4}{4! n^2} \sum_{j=1}^{n} V_j^4 + \frac{t^4}{2!2! n^2} \sum_{j_1=1}^{n} V_{j_1}^2 \sum_{j_2=1}^{n} V_{j_1}^2 + O(5).$$

For the sake of convenience, let us further introduce the average of the moments E_j^m:

$$\overline{V^m} = \frac{1}{n} \sum_{j=1}^{n} E_j^m. \tag{2.214}$$

Note the special case $\sigma^2 = \overline{V^2}$. In Eq. (2.213), we replace the sums by factors $n\overline{V^m}$. The characteristic function then reduces to

$$\phi_{z'}(t) = 1 - \frac{t^2}{2} \overline{V^2} - \frac{it^3}{3! n^{1/2}} \overline{V^3} + \frac{t^4}{4! n} \overline{V^4} + \frac{t^4}{2!2!} \overline{V^2 V^2} + O(5). \tag{2.215}$$

Terms in $1/n$ vanish in the large n limit. After substitution of $\overline{V^2}$ by σ^2, this expression thus further reduces to

$$\phi_{z'}(t) = 1 - \frac{t^2 \sigma^2}{2} + \frac{t^4 \sigma^4}{4} + O(6) = \exp(-\sigma^2 t^2/2), \tag{2.216}$$

which corresponds to the characteristic function of a Gaussian with mean zero and variance σ^2. Transforming back to the variable $\sum_i x_i$, one obtains a Gaussian with mean equal to $\sum_i \mu_i$ and variance $n\sigma^2 = \sum_i \sigma_i^2$. This completes the proof of the CLT.

A few words of caution are in order. First, the preceding discussion is strictly valid only when moments of the PDFs $f(x_i)$ exist. Distributions with very long tails, such as the

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

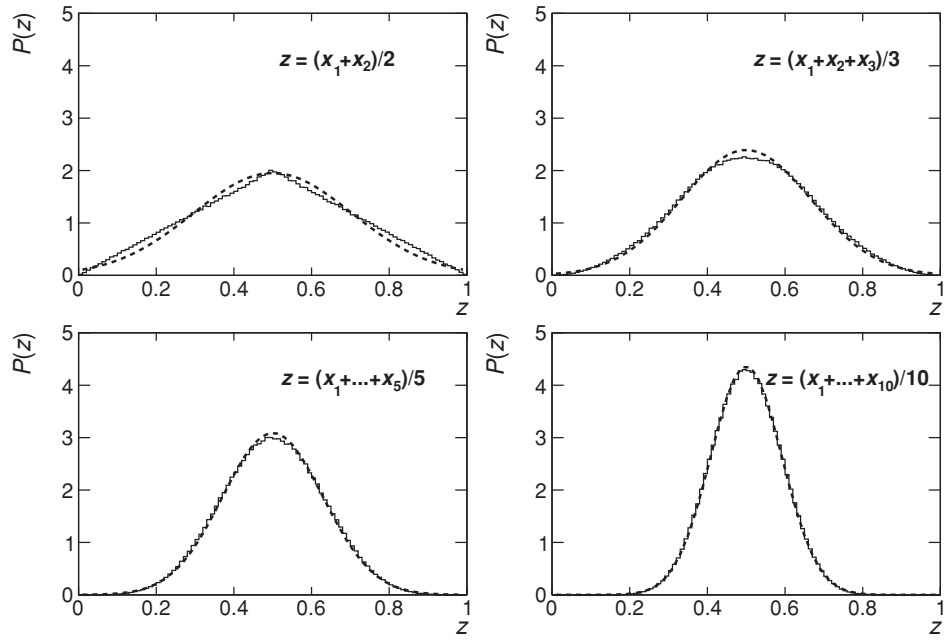

Fig. 2.12 Illustration of the Central Limit Theorem: Histograms of scaled sums ($z = \frac{1}{n}\sum_{i=1}^{n} x_i$) of $n = 2, 3, 5$, and 10 uniformly distributed random deviates in the range [0, 1] obtained with 1 million entries. Dashed curves: Gaussian distribution with standard deviations $\sigma = 1/\sqrt{12n}$.

Breit–Wigner distribution or the Landau distribution,[9] which have divergent expectation values $\mathrm{E}[x^k]$, will thus elude the CLT. In practice, in specific analyses one may be more concerned with the rate at which the PDF of a finite sum of many random variables converges to a Gaussian distribution. Obviously, as per Eq. (2.198), a sum of Gaussian deviates itself follows a Gaussian distribution. It is, however, difficult to quantify the rate at which the sum of variables with arbitrary PDFs might converge to a Gaussian distribution without a sophisticated and detailed analysis. Still, it is usually relatively easy to test whether sufficient convergence is achieved at given n by using simple Monte Carlo simulations such as those illustrated in Figure 2.12 in which scaled sums of n uniformly distributed deviates, $z = \frac{1}{n}\sum_{i=1}^{n} x_i$, are compared with a normal distribution of mean $\mu = 0.5$ and standard deviation $\sigma = 1/\sqrt{12n}$.

2.11 Measurement Errors

Consider repeated measurements of a set of m observables $\vec{X} = (X_1, X_2, \ldots, X_m)$ yielding a set of n random variables $\{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n\}$. If the experimental conditions of the n

[9] A probability distribution named after Soviet physicist Lev Landau and commonly used to model the energy of particles traversing a medium of finite thickness.

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

measurements are identical, the dispersion of values is a result of the measurement process and it is reasonable to assume that the measured values are drawn from a common parent population with a PDF $f(\vec{x})$. The PDF $f(\vec{x})$ is evidently not completely known but one can use the measured sample to obtain estimates of the means $\vec{\mu} = (\mu_1, \mu_2, \ldots, \mu_m)$ and covariance matrix V_{ij} of the variables x_i. Techniques to obtain such estimates will be discussed in detail in Chapter 4. In this section, let us simply assume these estimates are somehow available.

2.11.1 Error Propagation

Suppose we wish to evaluate a certain observable Y based on a function $y(\vec{x})$ of the variables X_i. If the PDF $f(\vec{x})$ was known, one could proceed, as in §2.8.6, and obtain the PDF $g(y)$ corresponding to the function $y(\vec{x})$. With this PDF in hand, it would then be possible to evaluate the mean value μ_y, which one could then adopt as the known value of the observable Y. The standard deviation σ_y would then characterize the uncertainty on μ_y. But, since $f(\vec{x})$ is unknown, it is not possible to determine $g(y)$ ab initio. What can then be done?

It turns out that although it is not possible to fully determine the PDF $g(y)$, one can nonetheless estimate its mean μ_y and variance σ_y based solely on the means $\vec{\mu}$ and the covariance matrix V_{ij}. This can be accomplished by calculating the Taylor expansion of the function $f(\vec{x})$ in the vicinity of $\vec{x} = \vec{\mu}$ truncated to first order:

$$y(\vec{x}) = y(\vec{\mu}) + \sum_{i=1}^{n} \left[\frac{\partial y}{\partial x_i}\right]_{\vec{x}=\vec{\mu}} (x_i - \mu_i) + O(2). \tag{2.217}$$

By definition, one has $\mathrm{E}[x_i - \mu_i] = 0$, and the expectation value of y is thus to first order:

$$\mathrm{E}[y(\vec{x})] \approx y(\vec{\mu}). \tag{2.218}$$

Let us next calculate the expectation value of y^2 to estimate the variance of the preceding estimate:

$$\mathrm{E}[y^2(\vec{x})] = y^2(\vec{\mu}) + 2y(\vec{\mu}) \sum_{i=1}^{n} \left[\frac{\partial y}{\partial x_i}\right]_{\vec{\mu}} E\left[x_i - \mu_i\right] \tag{2.219}$$

$$+ E\left[\left(\sum_{i=1}^{n} \left[\frac{\partial y}{\partial x_i}\right]_{\vec{\mu}} (x_i - \mu_i)\right)\left(\sum_{j=1}^{n} \left[\frac{\partial y}{\partial x_j}\right]_{\vec{\mu}} (x_j - \mu_j)\right)\right] + O(3).$$

The second term cancels out because $\mathrm{E}[x_i - \mu_i] = 0$, and we must therefore keep the next order in the Taylor expansion. Given the derivatives $\partial y/\partial x_i|_{\vec{x}=\vec{\mu}}$ are constants, the third term can be rewritten

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \left(\left[\frac{\partial y}{\partial x_j}\right]_{\vec{x}=\vec{\mu}} \left[\frac{\partial y}{\partial x_i}\right]_{\vec{x}=\vec{\mu}} E\left[(x_j - \mu_j)(x_i - \mu_i)\right]\right). \tag{2.220}$$

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

The expectation value $E\left[\left(x_j - \mu_j\right)\left(x_i - \mu_i\right)\right]$ is the covariance of the variables x_i and x_j, which we denote V_{ij}. The expectation value $\mathrm{E}[y^2(\vec{x})]$ may thus be written

$$\mathrm{E}[y^2(\vec{x})] \approx y^2(\vec{\mu}) + \sum_{i,j=1}^{n} \left[\frac{\partial y}{\partial x_i}\right]_{\vec{x}=\vec{\mu}} \left[\frac{\partial y}{\partial x_j}\right]_{\vec{x}=\vec{\mu}} V_{ij}, \tag{2.221}$$

from which we conclude that the variance σ_y^2 is given to first order by

$$\sigma_y^2 \approx \sum_{i,j=1}^{n} \left[\frac{\partial y}{\partial x_i}\frac{\partial y}{\partial x_j}\right]_{\vec{x}=\vec{\mu}} V_{ij}. \tag{2.222}$$

More generally, given a set of function of m functions $y_1(\vec{x}), \cdots, y_m(\vec{x})$, the covariance matrix of these variables may be calculated as follows (see Problem 2.24):

$$U_{kl} = \mathrm{Cov}[y_k, y_l] \approx \sum_{i,j=1}^{n} \left[\frac{\partial y_k}{\partial x_i}\frac{\partial y_l}{\partial x_j}\right]_{\vec{x}=\vec{\mu}} V_{ij}. \tag{2.223}$$

Given that the quantities involved have finitely many indices, $k, l = 1, \ldots, m$ and $i, j = 1, \ldots, n$, it is convenient to represent Eq. (2.223) in matrix notation as follows:

$$\mathbf{U} = \mathbf{A}\mathbf{V}\mathbf{A}^T, \tag{2.224}$$

where we introduced a matrix of derivatives $\mathbf{A}$ defined as

$$A_{ij} = \left.\frac{\partial y_i}{\partial x_j}\right|_{\vec{x}=\vec{\mu}}, \tag{2.225}$$

and the notation $\mathbf{A}^T$ stands for the transpose of matrix $\mathbf{A}$:

$$\left(\mathbf{A}^T\right)_{ij} = (\mathbf{A})_{ji}. \tag{2.226}$$

Equations (2.222, 2.223) provide the basis for error propagation used in multivariate scientific problems. They are also used in signal processing, for example, with Kalman filtering discussed in §5.6.

2.11.2 Uncorrelated Error Propagation

Error propagation readily simplifies when the n variables x_i are mutually independent, that is, uncorrelated. In such cases, one has $V_{ii} = \sigma_i^2$ and $V_{ij} = 0$ for $i \neq j$. Eqs. (2.222, 2.223) thus reduce to

$$\sigma_y^2 = \sum_{i=1}^{n} \left[\frac{\partial y}{\partial x_i}\right]_{\vec{x}=\vec{\mu}}^2 \sigma_i^2, \tag{2.227}$$

and

$$U_{kl} = \sum_{i=1}^{n} \left[\frac{\partial y_k}{\partial x_i}\frac{\partial y_l}{\partial x_i}\right]_{\vec{x}=\vec{\mu}} \sigma_i^2. \tag{2.228}$$

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

It is important to note that the partial derivatives $\partial y_k/\partial x_i$ and $\partial y_l/\partial x_i$ are generally nonzero, the off-diagonal matrix elements U_{kl}, $k \neq l$ are thus nonvanishing, and the matrix $\mathbf{U}$ is nondiagonal even though the covariance matrix $\mathbf{V}$ is.

2.11.3 Basic Examples of Error Propagation

Equations (2.222–2.228) may be used, for instance, to estimate the errors on sums and products of random variables. For a sum $y = x_1 + x_2$, one finds the error is (see Problem 2.25)

$$\sigma_y^2 \approx \sigma_1^2 + \sigma_2^2 + 2V_{12} \quad (\text{error on} \quad y = x_1 + x_2), \tag{2.229}$$

whereas for a product $y = x_1 \times x_2$, one gets (see Problem 2.27)

$$\frac{\sigma_y^2}{y^2} \approx \frac{\sigma_1^2}{x_1^2} + \frac{\sigma_2^2}{x_2^2} + \frac{2V_{12}}{x_1 x_2} \quad (\text{error on} \quad y = x_1 \times x_2). \tag{2.230}$$

The preceding expressions simplify to the *usual* equations introduced in elementary physics courses when variables x and y are independent, i.e., when $V_{12} = 0$.

Note that, as per our discussion of the sum of Gaussian deviates in §2.10.3, the estimate given by Eq. (2.229) of the variance of a sum $y = x_1 + x_2$ is exact if the variables x_1 and x_2 are Gaussian deviates.

2.11.4 Covariance Matrix Diagonalization

Consider once again a large set of joint measurements of m random variables $x_1, \ldots, x_m$ yielding mean values $\mu_1, \ldots, \mu_m$ and a covariance matrix $V_{ij} = \text{Cov}[x_i, x_j]$. Nonvanishing off-diagonal elements of this matrix shall indicate the random variables are correlated and thus not independent of one another. Characterizing errors on such a system of variables is nontrivial in general. It should indeed be clear that it is factually incorrect to use only the variances V_{ii} to state uncertainty estimates for the variables x_i. Indeed, since the x_i are mutually correlated, so are their errors. We will see in §6.1.5 that statements about the uncertainties of correlated variables require correlated error regions, which may be quite cumbersome to represent for $m > 2$. However, should there be a way to obtain a system of m variables $y_1, \ldots, y_m$ by linear transformation of the random variables x_i

$$y_i = \sum_{j=1}^{m} R_{ij} x_j, \tag{2.231}$$

such that the covariance matrix $U_{ij} = \text{Cov}[y_i, y_j]$ is diagonal, the characterization and modeling of the system could be carried out in terms of m independent variables and would thus be far simpler.

Let us verify that the preceding linear transformation exists by inserting Eq. (2.231) for y_i in the formula of the covariance matrix elements U_{ij}:

$$U_{ij} = \sum_{k,k'=1}^{m} R_{ik} R_{jk'} \text{Cov}\,[x_k, x_{k'}]. \tag{2.232}$$

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

Substituting $V_{kk'}$ for Cov $[x_k, x_{k'}]$, we get

$$U_{ij} = \sum_{k,k'=1}^{m} R_{ik} V_{kk'} R_{jk'} \tag{2.233}$$

$$= \sum_{k,k'=1}^{m} R_{ik} V_{kk'} R^T_{k'j}, \tag{2.234}$$

which in matrix notation may be written

$$\mathbf{U} = \mathbf{R}\mathbf{V}\mathbf{R}^T. \tag{2.235}$$

In order to identify the elements of the matrix **R**, let us consider the linear equation

$$V\vec{r}^{(i)} = \alpha_i \vec{r}^{(i)}, \tag{2.236}$$

where α_i and $\vec{r}^{(i)}$ are eigenvalues and eigenvectors of **V**, respectively. Techniques to determine eigenvalues are presented in various textbooks on linear algebra and are thus not discussed here. By construction, the eigenvectors are required to obey the orthogonality condition

$$\vec{r}^{(i)} \cdot \vec{r}^{(j)} = \sum_{k=1}^{m} r_k^{(i)} r_k^{(j)} = \delta_{ij}. \tag{2.237}$$

In cases where two or more eigenvalues are equal, the direction of the corresponding vectors are not uniquely defined but can be chosen arbitrarily to meet the foregoing condition. The n rows of the transformation matrix **R** may then be written as the n eigenvectors $\vec{r}^{(i)}$ of the matrix **V**, and one can verify by simple substitution that **U** has the required property, that is, it consists of a diagonal matrix:

$$U_{ij} = \sum_{k,l=1}^{m} R_{ik} V_{kl} R^T_{lj} = \sum_{k,l=1}^{m} r_k^{(i)} V_{kl} r_l^{(j)}, \tag{2.238}$$

$$= \alpha_j \sum_{k=1}^{m} r_k^{(i)} r_k^{(j)}, \tag{2.239}$$

$$= \alpha_j \delta_{ij}. \tag{2.240}$$

We conclude that the variances of the transformed variables $y_1, \ldots, y_n$ are given by the eigenvalues α_i of the original covariance matrix V, and indeed that all off-diagonal elements of the covariance matrix **U** are null. It is thus possible to model or characterize the system of variables $\{x_i\}$ on the basis of the n independent random variables y_i.

Note in closing that by virtue of Eq. (2.237), one finds that the transpose of matrix **R** is equal to its inverse, $\mathbf{R}^{-1} = \mathbf{R}^T$ (see Problem 2.32). The matrix **R** consequently corresponds to an orthogonal transformation carrying a rotation of the vector x that leaves its norm invariant.

Analytical solutions of Eq. (2.236) are simple for 2×2 and 3×3 matrices but rapidly become cumbersome or even intractable for larger matrices. Problem 2.33 discusses a case involving the diagonalization of a 2×2 covariance matrix analytically. Diagonalization of

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

larger matrices are usually carried out numerically using popular software packages such as MATLAB®, Mathematica®, and ROOT [59].

2.12 Random Walk Processes

A random walk is a process consisting of a succession of random steps. It could be the path followed by a molecule as it travels in a liquid or a gas, the behavior of financial markets, or the production of particles with collective behavior (e.g., flow). As it turns out, the notion of random walk is central to the description of many stochastic phenomena, most notably diffusion processes and the description of collective flow phenomena observed in heavy-ion collisions. Diffusion and scattering processes amounts to a succession (i.e., a sum) of many "small" individual scattering processes. They thus constitute a natural application of the central limit theorem.

We here discuss random walk processes of increasing complexity starting, in §2.12.1, with random motion in one dimension. The discussion is then extended to two dimensions in §2.12.2. Extensions to three or more dimensions are readily possible and addressed in the problem section.

2.12.1 Random Walks in One Dimension

Let us first consider a random walk in one dimension with fixed step size along the x-axis. Starting from the origin, a walker flips a fair coin to decide in which direction to go next. Heads, he moves along $+x$ by one unit; tails, he moves along $-x$ by one unit. After he lands at the new position, he flips the coin, and repeats the process for several steps denoted by an index i. Each move is represented by $x_i = \pm 1$ and the two values have equal probability: $P(+1) = P(-1) = 0.5$.

To determine the probability of finding the walker at certain position x after a large number of steps n, let us consider the sum of all steps from the walker's initial position, $S_n = \sum_{j=1}^{n} x_j$. Our task is thus to determine the probability density of finding the walker at certain position S_n. The series $\{S_n\}$ is called simple random walk on the set of integers $\mathbf{Z}$.

Let us first evaluate the mean and variance of S_n. Since the values $x = +1$ and $x = -1$ are equally probable, the expectation value of x (a single step) is $\langle x \rangle = \mathrm{E}[x] = 0$ and its variance equal $\sigma_1 = \mathrm{Var}[x] = 1$. The mean and variance of S_n are thus, respectively,

$$\langle S_n \rangle = \mathrm{E}[S_n] = \sum_{j=1}^{n} \mathrm{E}[x_j] = 0, \tag{2.241}$$

$$\sigma_n^2 = \mathrm{Var}[S_n] = \sum_{j=1}^{n} \mathrm{E}[x_j^2] + \sum_{i \neq k=1}^{n} \mathrm{E}[x_i x_j] = n, \tag{2.242}$$

since $\mathrm{E}[x_i x_j] = \mathrm{E}[x_i]\mathrm{E}[x_j] = 0$ for all $i \neq j$, and $\mathrm{E}[x_j^2] = 1$.

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

To find the PDF of S_n, we invoke the central limit theorem, and obtain in the large n limit,

$$\frac{1}{N}\frac{dN}{dS_n} = \frac{1}{\sqrt{2\pi}\sigma_n}\exp\left(-\frac{S_n^2}{2\sigma_n^2}\right). \tag{2.243}$$

The foregoing reasoning can be repeated for a one-dimensional random walk with variable step size. Assuming for instance that the step size is uniformly distributed in the range $[-L/2, L/2]$, the expectation value of a single step remains null, and its variance is $\sigma_1^2 = L^2/12$, as per Eq. (3.89). The expectation value of the sum S_n thus also remains null in the large n limit, and its variance becomes $\sigma_n^2 = n\sigma_1^2$, while the form of the PDF remains the same.

It is also interesting to consider the introduction of a biased random step. Returning for instance to the case of a fixed size step random walk, let us assume that the probability of a step in the positive direction is p while that of a step in the negative direction is $1-p$. The expectation value of a single step then becomes $\mathrm{E}[x] = 2p-1$ and its variance $\mathrm{Var}[x] = 2p(1-p) = \sigma_1^2$. The mean of the sum is consequently shifted, $\langle S_n\rangle = n(2p-1)$, and the variance becomes $\sigma_n^2 = 2np(1-p) = n\sigma_1^2$. The PDF of S_n is then

$$\frac{1}{N}\frac{dN}{dS_n} = \frac{1}{\sqrt{2\pi}\sigma_n}\exp\left(-\frac{(S_n-\langle S_n\rangle)^2}{2\sigma_n^2}\right). \tag{2.244}$$

Alternatively, one may also consider a forward biased Gaussian distributed random walk. Let the PDF of one step, x, be given by

$$\frac{1}{N}\frac{dN}{dx} = \frac{1}{\sqrt{2\pi}\sigma_1}\exp\left(-\frac{(x-\langle x\rangle)^2}{2\sigma_1^2}\right), \tag{2.245}$$

where $\langle x\rangle$ and σ_1 are respectively the mean and RMS of the step size. By virtue of the addition theorem (2.193), the PDF of the sum S_n of n random steps thus has the same form as Eq. (2.244) but with mean $\langle S_n\rangle = n\langle x\rangle$ and variance $\sigma_n^2 = n\sigma_1^2$.

2.12.2 Random Walks in 2D

We next proceed to determine the characteristics of a two-dimensional random walk with fixed step but arbitrary direction. Let $x_i = \cos\phi_i$ and $y_i = \sin\phi_i$ be the projections of the unit step size in the the x–y plane along the x- and y-axes, respectively. All directions ϕ_i being equally probable, we write $P(\phi_i) = (2\pi)^{-1}$ in the range $[0, 2\pi]$. Sums of projections along the x- and y-axes are denoted $S_{x,n} = \sum_{j=1}^n x_i$ and $S_{y,n} = \sum_{j=1}^n y_i$, respectively. They define a displacement vector $\vec{S}_n$, with modulus $S_n = \sqrt{S_{x,n}^2 + S_{y,n}^2}$, and direction relative to the x-axis given by $\psi_n = \tan^{-1}(S_{y,n}/S_{x,n})$.

By construction, $\mathrm{E}[x_i] = \mathrm{E}[y_i] = 0$, $\mathrm{E}[x_i^2] = \mathrm{E}[y_i^2] = \mathrm{E}[\cos^2\phi_i] = 1/2$, which entails $\mathrm{E}[S_{x,n}] = \mathrm{E}[S_{y,n}] = 0$ and $\sigma_n^2 = \mathrm{Var}[S_{x,n}] = \mathrm{Var}[S_{y,n}] = n/2$. In the large n limit, both $S_{x,n}$ and $S_{y,n}$ are Gaussian distributed:

$$\frac{1}{N}\frac{d^2N}{dS_{x,n}dS_{y,n}} = \frac{1}{2\pi\sigma_n^2}\exp\left(-\frac{S_{x,n}^2}{2\sigma_n^2}\right)\exp\left(-\frac{S_{y,n}^2}{2\sigma_n^2}\right), \tag{2.246}$$

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

which simplifies to

$$\frac{1}{N}\frac{d^2N}{S_n dS_n d\psi} = \frac{1}{2\pi\sigma_n^2}\exp\left(-\frac{S_n^2}{2\sigma_n^2}\right) \tag{2.247}$$

after substitution of the modulus of $\vec{S}_n$ for the sum of the square of its components. Note that the PDF is independent of ψ. One thus concludes, as expected, that all directions are equally probable. Integration over the angle ψ yields

$$\frac{1}{N}\frac{dN}{dS_n} = \frac{S_n}{\sigma_n^2}\exp\left(-\frac{S_n^2}{2\sigma_n^2}\right). \tag{2.248}$$

The variance $\sigma_n^2 = n\sigma_1$ implies the "typical" length of S_n scales as the square root of the number of steps, that is, $\sqrt{n}$.

The foregoing two-dimensional random walk can be modified to describe motion biased toward an arbitrary direction ψ. This type of biased random walk can be used to model collective flow, that is, the concerted motion of produced particles, observed in the study of heavy-ion collisions at medium to high energy. To this end, let $P(\phi_i) = (2\pi)^{-1}[1 + 2v_m\cos(m(\phi_i - \psi))]$, where the coefficient v_m, commonly called **flow coefficient of order** m is usually much smaller than unity. For a fixed value of ψ, and a fixed step size, r, the expectation values of $x_i = r\cos(m\phi_i)$ and $y_i = r\sin(m\phi_i)$ are (see Problem 2.35):

$$\mathrm{E}[x_i] = v_m r\cos(m\psi), \tag{2.249}$$

$$\mathrm{E}[y_i] = v_m r\sin(m\psi), \tag{2.250}$$

while the second moments are

$$\mathrm{E}[x_i^2] = \mathrm{E}[y_i^2] = \frac{r^2}{4\pi}. \tag{2.251}$$

The variances of x_i and y_i are thus respectively

$$\sigma_{x,1} \equiv \mathrm{Var}[x_i] = \frac{r^2}{4\pi} - r^2 v_m^2\cos^2(m\psi) \approx \frac{r^2}{4\pi}, \tag{2.252}$$

$$\sigma_{y,1} \equiv \mathrm{Var}[y_i] = \frac{r^2}{4\pi} - r^2 v_m^2\sin^2(m\psi) \approx \frac{r^2}{4\pi}, \tag{2.253}$$

where the right-hand side approximations hold if the coefficients are small ($v_m << 1$). The expectation values of the sums $S_{x,n}$ and $S_{y,n}$ are consequently

$$\langle S_{x,n}\rangle = \mathrm{E}[S_{x,n}] = nrv_m\cos(m\psi), \tag{2.254}$$

$$\langle S_{y,n}\rangle = \mathrm{E}[S_{y,n}] = nrv_m\sin(m\psi), \tag{2.255}$$

while their standard deviations are

$$\sigma_{x,n} = \sqrt{n}\sigma_{x,1}, \tag{2.256}$$

$$\sigma_{y,n} = \sqrt{n}\sigma_{y,1}. \tag{2.257}$$

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

In the large n limit, both $S_{x,n}$ and $S_{y,n}$ are Gaussians

$$\frac{1}{N}\frac{dN}{dS_{x,n}} = \frac{1}{\sqrt{2\pi}\sigma_{x,n}} \exp\left[-\frac{(S_{x,n} - \langle S_{x,n}\rangle)^2}{2\sigma_{x,n}^2}\right], \tag{2.258}$$

$$\frac{1}{N}\frac{dN}{dS_{y,n}} = \frac{1}{\sqrt{2\pi}\sigma_{y,n}} \exp\left[-\frac{\left(S_{y,n} - \langle S_{y,n}\rangle\right)^2}{2\sigma_{y,n}^2}\right]. \tag{2.259}$$

For small values of v_m, one has $\sigma_{x,n}^2 \approx \sigma_{y,n}^2 = \sigma_n^2 = n\frac{r^2}{4\pi}$, and one can thus write (see Problem 2.36)

$$\frac{1}{N}\frac{dN}{S_n dS_n d\theta} = \frac{1}{2\pi\sigma_n^2} \exp\left[-\frac{\left(\vec{S_n} - \langle\vec{S_n}\rangle\right)^2}{2\sigma_n^2}\right], \tag{2.260}$$

where we introduced $\vec{S_n} = (S_{x,n}, S_{y,n})$, its average

$$\langle\vec{S_n}\rangle = (nrv_m \cos(m\psi), nrv_m \sin(m\psi)), \tag{2.261}$$

and θ, the angle between the vectors $\vec{S_n}$ and $\langle\vec{S_n}\rangle$.

The vectors $\langle\vec{S_n}\rangle$ and $\vec{S_n}$ represent the expectation value and a particular realization of the random walk with parameter v_m, respectively. A single realization of the random walk, $\vec{S_n}$, can be regarded as a measurement of $\langle\vec{S_n}\rangle$. It is thus interesting to consider how precise the "measurements" of the modulus S_n and angle ψ are relative to the expectation values $\langle S_n\rangle$ and $\tan^{-1}(\langle S_{y,n}\rangle/\langle S_{x,n}\rangle)$. Integration of (2.260) over S_n yields (see Problem 2.37)

$$\frac{1}{N}\frac{dN}{d\theta} = \frac{1}{\pi}\exp(-\chi^2)\left\{1 + z\sqrt{\pi}\,[1 + \mathrm{erf}(z)]\exp(z^2)\right\} \tag{2.262}$$

where $z = \chi\cos(\theta)$, $\mathrm{erf}(x)$ is the error function, and $\chi \equiv \langle S_n\rangle/\sigma_n$.

Integrations over θ gives (see Problem 2.37)

$$\frac{1}{N}\frac{dN}{dS_n} = \frac{1}{\sigma_n^2}\exp\left(-\frac{(S_n^2 + \langle S_n\rangle^2)}{2\sigma_n^2}\right) I_0\left(\frac{\langle S_n\rangle|S_n|}{\sigma_n^2}\right), \tag{2.263}$$

where $I_0(z)$ represents the modified Bessel function of the first kind and of order 0. Note that the expressions (2.262) and (2.263) are independent of the step size r but depend on the magnitude of the coefficient v_m.

2.13 Cumulants

Cumulants provide a powerful tool for the study of multiple variable correlations and are the basis of several analysis techniques used in nuclear and particle physics, many of which are presented in Chapters 10 and 11.

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

2.13.1 Cumulant Definition

The cumulants κ_n of a random variable x, relative to a specific probability distribution, are formally defined in terms of the cumulant-generating function $g(t)$, which is the logarithm of the moment-generating function of that probability distribution:

$$g(t) = \ln\left(\mathrm{E}\left[e^{tx}\right]\right). \tag{2.264}$$

The cumulants κ_n are obtained as coefficients of the power series expansion of $g(t)$:

$$g(t) = \sum_{n=1}^{\infty} \kappa_n \frac{t^n}{n!}. \tag{2.265}$$

If $g(t)$ is available in close analytical form, the cumulants can then be calculated according to

$$\kappa_n = \left.\frac{d^n g(t)}{dt^n}\right|_{t=0} \tag{2.266}$$

As an example, consider the calculation of the cumulants of the binomial distribution with moment-generating function (from Table 3.1)

$$M(t) = \left(1 - p + pe^t\right)^n. \tag{2.267}$$

The function $g(t)$ becomes

$$g(t) = \ln\left(M(t)\right) = n \ln\left(1 - p + pe^t\right). \tag{2.268}$$

Derivatives of order n of $g(t)$ yield the cumulants

$$\kappa_1 = \left.\frac{d}{dt} g(t)\right|_{t=0} = np, \tag{2.269}$$

$$\kappa_2 = \left.\frac{d^2}{dt^2} g(t)\right|_{t=0} = np(1-p), \tag{2.270}$$

$$\kappa_3 = \left.\frac{d^3}{dt^3} g(t)\right|_{t=0} = n(2p^3 - 3p^2 + p), \tag{2.271}$$

etc.

The use of cumulants is convenient in the analysis of independent random variables consisting of a sum of two or more independent variables. For instance, let us define a random variable $z = x + y$, where x and y are two statistically independent variables with cumulant-generating functions $g_x(t)$ and $g_y(t)$, respectively. Let us calculate the cumulant-generating function $g_z(t)$ and show that cumulants of z of all orders n are equal to the sum of the n order cumulants of x and y (additivity property).

$$g_z(t) = \ln\left(\mathrm{E}\left[e^{t(x+y)}\right]\right), \tag{2.272}$$

$$= \ln\left(\mathrm{E}\left[e^{tx}\right]\mathrm{E}\left[e^{ty}\right]\right), \tag{2.273}$$

$$= \ln\left(\mathrm{E}\left[e^{tx}\right]\right) + \ln\left(\mathrm{E}\left[e^{ty}\right]\right), \tag{2.274}$$

$$= g_x(t) + g_y(t), \tag{2.275}$$

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

where in the second line we have use the statistical independence of x and y to factors their expectation values. Since the derivative (at any order) of a sum of two functions is equal to the sum of the derivatives of the functions (also at all orders), one finds that the cumulants of z equal the sum of the cumulants of x and y as follows:

$$\kappa_n^{(z)} = \left.\frac{d^n}{dt^n} g_z(t)\right|_{t=0}, \tag{2.276}$$

$$= \left.\frac{d^n}{dt^n} g_x(t)\right|_{t=0} + \left.\frac{d^n}{dt^n} g_y(t)\right|_{t=0}, \tag{2.277}$$

$$= \kappa_n^{(x)} + \kappa_n^{(y)}. \tag{2.278}$$

Let $c \in \mathbf{R}$, a constant. Cumulants are easily verified to have the following basic properties:

$$\kappa_1(x + c) = \kappa_1(x) + c \quad \text{shift} - \text{equivariance}, \tag{2.279}$$

$$\kappa_n(x + c) = \kappa_n(x) \qquad \text{for } n \geq 2, \text{ shift invariance}, \tag{2.280}$$

$$\kappa_n(cx) = c^n \kappa_n(x) \qquad \text{homogeneity}. \tag{2.281}$$

Additionally, note that by definition (2.264) of the cumulant-generating function $g(t)$, one can write

$$M(t) = \exp\left(g(t)\right), \tag{2.282}$$

$$1 + \sum_{n=1}^{\infty} \frac{\mu_n{}' t^n}{n!} = \exp\left(\sum_{k=1}^{\infty} \frac{\kappa_n t^n}{n!}\right). \tag{2.283}$$

Using this expression, one can derive the following recursion formula between the moments $\mu_n{}'$ and the cumulants κ_n,

$$\kappa_n = \mu_n{}' - \sum_{m=1}^{n-1} \binom{n-1}{m-1} \kappa_m \mu_{n-m}{}' \tag{2.284}$$

One finds that the first moment equals the first cumulant while the second and third central moments equal the second and third central cumulants, respectively. The fourth cumulant is related to the excess kurtosis, $\kappa_4 = \mu_4 - 3\mu_2^2$. Higher cumulants are more complicated polynomial functions of the moments.

2.13.2 Joint Cumulants

As for moments, one can also define joint cumulants. The joint cumulants of random variables $x_1, x_2, \ldots x_n$ are defined as derivatives of the joint cumulant-generating function

$$g(t_1, t_2, \ldots, t_n) = \ln\left(\mathrm{E}\left[e^{\sum_{j=1}^{n} t_j x_j}\right]\right). \tag{2.285}$$

The first-order cumulant of n random variables may be written

$$\kappa[x_1, x_2, \ldots, x_n] = \sum_{P} (|P| - 1)! \, (-1)^{|P|-1} \prod_{B \in P} \mathrm{E}\left[\prod_{i \in B} x_i\right], \tag{2.286}$$

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

where P stands for partitions of $\{1, 2, \ldots, n\}$, $|P|$ is the number of parts in such partitions, B runs through all the blocks of the partition, while i enumerates elements of any given partition.[10] With two variables x and y, one gets the covariance

$$\kappa[x, y] = \mathrm{E}[xy] - \mathrm{E}[x]\,\mathrm{E}[y]\,, \tag{2.287}$$

and with three variables x, y, and z, one obtains

$$\kappa[x, y, z] = \mathrm{E}[xyz] - \mathrm{E}[xy]\,\mathrm{E}[z] - \mathrm{E}[xz]\,\mathrm{E}[y] - \mathrm{E}[yz]\,\mathrm{E}[x] + 2\mathrm{E}[x]\,\mathrm{E}[y]\,\mathrm{E}[z]\,, \tag{2.288}$$

which for two identical variables, $x = y$, reduces to

$$\kappa[x, x, z] = \mathrm{E}\left[x^2 z\right] - \mathrm{E}[xz]\,\mathrm{E}[x] - \mathrm{E}\left[x^2\right]\mathrm{E}[z] + 2\mathrm{E}[x]^2\,\mathrm{E}[z]\,. \tag{2.289}$$

Given the expectation value of a product of statistically independent variables is the product of their expectation values, one easily verifies that any cumulant involving two or more independent variables is null by definition. And if all n random variables are identical, the joint cumulant is simply the nth ordinary cumulant.

The formula (2.286) expressing cumulants in terms of moments can be inverted. One finds

$$\mathrm{E}[x_1, x_2, \ldots, x_n] = \sum_P \prod_{B \in P} \kappa[x_i; i \in B]. \tag{2.290}$$

For instance, the moment of the product xyz is given by

$$\mathrm{E}[xyz] = \kappa[x, y, z] + \kappa[x, y]\kappa[z] + \kappa[x, z]\kappa[y] + \kappa[y, z]\kappa[x] + \kappa[x]\kappa[y]\kappa[z]. \tag{2.291}$$

Joint cumulants are quite important in the analysis of particle correlations in high-energy physics, most particularly in the study of multiparticle correlation functions and collective motion (flow) discussed in Chapters 10 and 11.

Exercises

2.1 Demonstrate the properties listed under Eq. (2.5).

2.2 Verify that the notion of conditional probability satisfies the three axioms defining the notion of probability.

2.3 Verify the expression Eq. (2.18) known as the law of total probability.

2.4 Derive the expression for the variance given by Eq. (2.80).

2.5 Show that the skewness, γ_1, can be equivalently defined as the ratio of the third cumulant κ_3 and the third power of the square root of the second cumulant $\kappa_2^{3/2}$.

2.6 Show that the excess kurtosis of a sum of n independent random variables x_i is equal to the sum of the kurtosis of these n variables divided by n^2 (Eq. 2.93).

[10] Examples of applications of the notions of cumulant, partitions, and blocks are presented in Chapter 11.

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

2.7 Derive the expression Eq. 2.60 for the density $g(q)$ obtained for a function $q(x)$ of continuous random variable x distributed according to a PDF, $f(x)$.

2.8 Calculate expressions for the density $g(a)\,da$ given functions $a(x) = x$ and $a(x) = x^4$, assuming the continuous variable x has a PDF, $f(x)$.

2.9 Verify that the full width at half maximum (FWHM) of a normal distribution is 2.35σ.

2.10 Derive a method to estimate the values $x_{i,o}$ defined by Eq. (4.64) for PDFs $f(x) \propto \exp(-x/\lambda)$ and $f(x) \propto (k+x)^{-\beta}$, where k and β are two unknown constants. Hint: Use interpolation of the yields in bins $i-1$ and $i+1$ to estimate the constants λ and β bin by bin.

2.11 Derive an expression similar to Eq. (2.118) for $h_{x_1,x_2}(x_1, x_2|x_3, \ldots x_m)$, where $m > 2$.

2.12 Show that the Pearson coefficients defined by Eq. (2.172) are bound in the range $-1 \leq \rho_{ij} \leq 1$ by construction.

2.13 Calculate the first, second, third, and fourth moments of the uniform distribution defined as follows:

$$p_U(x;\alpha,\beta) = \begin{cases} \frac{1}{\beta-\alpha} & \alpha \leq x \leq \beta \\ 0 & \text{otherwise} \end{cases} \tag{2.292}$$

2.14 Calculate the first, second, third, and fourth moments of the triangular distribution defined as follows:

$$p_T(x;\alpha,\beta) = \begin{cases} \frac{2(x-\alpha)}{(\beta-\alpha)^2} & \alpha \leq x \leq \beta \\ 0 & \text{otherwise} \end{cases} \tag{2.293}$$

2.15 Show that given a partition of a set S into n mutually disjoint subsets A_i, with $i = 1, n$, the probability $P(B)$, of a set $B \subset S$ may be written as follows:

$$P(B) = \sum_i P(B|A_i)P(A_i). \tag{2.294}$$

Hint: Disjoints subsets have a null intersection, $A_i \cap A_j = 0$ for $i \neq j$.

2.16 Combine Bayes' theorem with the law of total probability to obtain the following expression:

$$P(A|B) = \frac{P(B|A)P(A)}{\sum_i P(B|A_i)P(A_i)}. \tag{2.295}$$

2.17 A neural network is designed and trained to classify particles entering an electromagnetic calorimeter as photon (P), electron (E), or hadron (H) based on the longitudinal and transverse patterns of energy deposition they produce in the calorimeter. A Monte Carlo simulation is used to estimate the neural net performance summarized in Table 2.1. The notations P_a, E_a, and H_a are used to indicate that the energy deposition pattern is due to a photon, an electron, or a hadron. Data analyzed with the neural network provide for fractions 0.05, 0.15, and 0.8 of photons, electrons, and hadrons, respectively. Calculate the relative rates produced by the nuclear reaction under study.

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

Table 2.1 Particle Identification Probabilities Used in Problem 2.17

$P(P_a\|P) = 0.98$	$P(P_a\|E) = 0.06$	$P(P_a\|H) = 0.05$
$P(E_a\|P) = 0.01$	$P(E_a\|E) = 0.90$	$P(E_a\|H) = 0.15$
$P(H_a\|P) = 0.01$	$P(P_a\|E) = 0.04$	$P(H_a\|H) = 0.80$

2.18 Derive the expression (3.62) giving the moments of a PDF in terms of derivatives of its characteristic function.

2.19 Use Eq. (3.62) to calculate the moments of (a) the uniform distributions, (b) the exponential distribution, (c) the t-distributions, and (d) the χ^2-distribution.

2.20 Show that the difference between two independent Gaussian deviates $\Delta x = x_1 - x_2$ has a mean $\mu_{\Delta x} = \mu_1 - \mu_2$ and a variance $\sigma^2_{\Delta x} = \sqrt{\sigma_1^2 + \sigma_2^2}$.

2.21 Extend Eq. (2.229) and find an expression for the error of a sum of several independent Gaussian deviates $x_1, x_2, \ldots, x_n$.

2.22 Extend Eq. (2.229) and find an expression for the error of a sum of several independent Poisson deviates $x_1, x_2, \ldots, x_n$.

2.23 Show that the characteristic function of $f(w) = \frac{1}{\sqrt{2\pi w}} e^{-w/2}$ is given by $\phi_w(t) = (1 - 2it)^{-1/2}$.

2.24 Derive the expression (2.223) for the covariance of functions $y_i(\vec{x})$ and $y_j(\vec{x})$.

2.25 Demonstrate the expression $\sigma_y^2 \approx \sigma_1^2 + \sigma_2^2 + 2V_{12}$, given by Eq. (2.229), corresponding to the error on the sum of random variables x_1 and x_2.

2.26 Demonstrate the expression $\sigma_y^2 \approx \sigma_1^2 + \sigma_2^2 + 2V_{12}$, given by Eq. (2.229), corresponding to the error on the difference of random variables x_1 and x_2.

2.27 Demonstrate Eq. (2.230) corresponding to the error on the product of correlated random variables x_1 and x_2.

2.28 Calculate the error on a ratio $y = x_1/x_2$ obtained by dividing correlated random variables x_1 and x_2.

2.29 Show that $\lambda_\pm = \frac{1}{2}\left[\sigma_1^2 + \sigma_2^2 \pm \sqrt{\sigma_1^2 + \sigma_2^2 - 4\,(1-\rho^2)\,\sigma_1^2\sigma_2^2}\right]$ indeed satisfies Eq. (5.66), and find the eigenvectors r_+ and r_-.

2.30 Show that if the intersection of two subsets A and B is null (i.e., for $A \cap B = 0$), the subsets cannot be independent, and determine the value of $P(A \cap B)$.

2.31 Given PDFs $g(x)$ and $h(y)$ for random variables x and y respectively, calculate the PDF of variable z defined as

- $z^2 = x^2 + y^2$.
- $\tan^{-1}(x/y)$.

2.32 Verify by direct substitution of the eigenvectors $\vec{r}$ into $\mathbf{A}$ defined by Eq. (2.231) that $\mathbf{A}$ satisfies $\mathbf{A}^{-1} = \mathbf{A}^T$ (i.e., its inverse is equal to its transpose), and that it is consequently an orthogonal transformation of the vector x that leaves its norm invariant.

2.33 Consider a two-dimensional covariance matrix $\mathbf{V}$ defined as follows:

$$\mathbf{V} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}. \tag{2.296}$$

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press

Solve the eigenvalue equation $(\mathbf{V} - \lambda)\vec{r} = 0$ and show the eigenvalues may be written

$$\lambda_{\pm} = \frac{1}{2}\left[\sigma_1^2 + \sigma_2^2 \pm \sqrt{\sigma_1^2 + \sigma_2^2 - 4(1-\rho^2)\sigma_1^2\sigma_2^2}\right] \tag{2.297}$$

with eigenvectors given by

$$r_+ = \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix} r_- = \begin{pmatrix} -\sin\theta \\ \cos\theta \end{pmatrix} \tag{2.298}$$

such that

$$\theta = \frac{1}{2}\tan^{-1}\left(\frac{2\rho\sigma_1\sigma_2}{\sigma_1^2 - \sigma_2^2}\right)$$

and

$$\mathbf{A} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}.$$

2.34 Consider a system with two random (or fluctuating) observables x and y. Explore the types of correlation that might arise between these two observables by using a linear combination of two randomly generated numbers r_1 and r_2:

$$\begin{aligned} x &= a + b_1 r_1 + b_2 r_2 \\ y &= c + d_1 r_1 - d_2 r_2 \end{aligned} \tag{2.299}$$

where a, b_1, b_2, c, d_1, and d_2 are arbitrary constants. Calculate the mean and variance of the observable x and y as well as the covariance of x and y. Discuss conditions under which x and y might yield (a) independent variables, (b) correlated variables with a positive covariance, (c) correlated variables with a negative covariance, and (d) correlated variables with a Pearson coefficient equal to unity. Assume the random variables r_1 and r_2 have null expectation values, $E[r_1] = E[r_2] = 0$, and nonvanishing variances σ_1^2 and σ_2^2, respectively.

2.35 Verify that the first-order and second-order moments associated with a two-dimensional random walk are given by Eqs. (2.249) and (2.251).

2.36 Derive Eq. (2.260) from Eq. (2.258) for the expression of the probability density $\frac{1}{N}\frac{dN}{S_n dS_n d\theta}$ of the sum vector $\vec{S}_n$.

2.37 Verify the expressions (2.262) and (2.263) by explicitly integrating Eq. (2.260). Hints:

$$\begin{aligned} I_o(z) &= \frac{1}{\pi}\int_o^{\pi} e^{\pm z\cos(\theta)} d\theta \\ I_n(z) &= \frac{1}{\pi}\int_o^{\pi} e^{\pm z\cos(n\theta)} d\theta \end{aligned} \tag{2.300}$$

2.38 Show that the expression $\mathbf{U} = \mathbf{AVA}^T$ is equivalent to Eq. (2.223).

https://doi.org/10.1017/9781108241922.004 Published online by Cambridge University Press