2

Probability Theory

2.1 Introduction

By using the laws of classical mechanics, in principle we can make *exact* predictions of events by knowing the *exact* initial conditions (i.e., all positions and velocities of the relevant degrees of freedom). However, in practice, there are several events that are unpredictable, essentially because it is impossible to have the exact knowledge of the initial conditions and a very small error in those conditions will grow exponentially in time, invalidating any attempt to follow the exact equations of motion. When tossing a coin or rolling a die, we do not know the outcome of the event, but we can give some *probability* to each event, e.g., $1/2$ for head and $1/2$ for tail when tossing a coin, or $1/6$ for each side of a die (we assume that we are playing with fair coins and dice). The probability gives the measure of the likelihood that a given event will occur. Of course, it is based upon a mathematical approach, which transforms the unpredictability into something that is somehow predictable.

This idea seems very simple, but it took several hundred years to capture it. Indeed, since the ancient times, people in Greece and in the Roman Empire (but also in Asia, for example, in India) were tenacious gamblers; nevertheless, nobody tried to understand how the random events were related to mathematical laws. Many quarrels and disputes were resolved by tossing a coin, and the result was seen as the manifestation of the "celestial will." The human superstition represented a huge obstacle to define a scientific (i.e., mathematical) approach to random events. Eventually, after several hundreds of years, superstition was overcome by an even stronger human impulse: the desire of obtaining an economical profit.

The birth of the mathematical theory of probability is due to the studies done by Girolamo Cardano, who realized that for equiprobable events (like tossing a coin or rolling a die), the probability that a single event will appear is equal to 1 over the number of all possible events (independently from any celestial will). Therefore, the

https://doi.org/10.1017/9781316417041.003 Published online by Cambridge University Press

probability to obtain head when tossing a coin is 1/2 (equal to the one of obtaining tail); the probability of getting 1 when rolling a die is 1/6, and the probability of obtaining an odd number is 3/6. Cardano was very often short of money and kept himself solvent by being an accomplished gambler and chess player. His book *Liber de ludo aleae*, written around 1560 but not published until 1663, contains the first systematic treatment of probability.

Even though Cardano produced the first exposition of random events, the most important conceptual leap toward the modern theory of probability was given by Blaise Pascal and Pierre de Fermat in the 17th Century, when answering few questions posed by the Chevalier de Méré, who was a regular gambler in France. There, one popular game was to roll two dice several times with a bet on having at least one double 1 (ace). Chevalier de Méré wanted to know how many trials one must do to have a profitable game (i.e., that the probability of having a double ace will be larger than 1/2). Another question was a bit more complex and opened the real basis of the theory of probability. The problem can be illustrated in this simple way: there are two players (e.g., Fermat and Pascal), who are tossing a coin: head gives one point to Fermat, while tail gives one point to Pascal. The first of the two players who achieves 3 points will win a pot of 100 francs. However, for some reason, the two gamblers must interrupt the game at the point where Fermat has 2 points and Pascal only 1; how is the 100 franc pot to be divided?

Let us briefly discuss the solutions of these questions. As far as the first one is concerned, instead of computing directly the probability P of favorable events, which requires a cumbersome calculation, it is much easier to evaluate the probability Q of unfavorable events and then take its complement, i.e., $P = 1 - Q$. In this example, the probability of obtaining a double ace (the successful event) when rolling a die is 1/36, while the probability that this event does not appear is obviously 35/36. The probability to obtain at least one double ace in N trials is therefore $1 - (35/36)^N$, namely it is 1 minus the probability that no successful events appear in N trials. For $N = 25$, the probability of having at least a double ace is larger than 1/2, so it becomes profitable to bet on it. Regarding the second question, the central point is that we do not have to see what happened in the past but what could happen in the future. Indeed, the following two answers are not correct: (a) Fermat takes all because he won more than Pascal and (b) Fermat takes twice more than Pascal since the result is 2−1. The correct one is to see what are the possibilities if they could have continued the game. Then, the possible outcomes are: head-head, head-tail, tail-head, and tail-tail; only in the latter case, Pascal would have won, while in the other three cases Fermat would have gained. Therefore, Fermat must take 3/4 of the total pot and Pascal only 1/4. Today, this result may look obvious; however, in the seventeenth century the idea that future events may be treated in a rigorous mathematical way has represented an incredible step forward.

https://doi.org/10.1017/9781316417041.003 Published online by Cambridge University Press

There are several definitions for the probability, which are rooted in different philosophical approaches. The *classical definition* dates back to Pierre Simon Laplace, who defined the probability of a given circumstance as the ratio between the number of favorable events divided by the total number of possible events, provided all the events are equiprobable. This definition is not satisfactory since it requires the concept of equiprobability itself. Moreover, it does not give a definition when the events are not equiprobable and assumes a finite number of outcomes (and, therefore, it does not apply for continuous variables). To overcome these difficulties, the frequentist definition of probability was introduced: it defines the probability of an event as the limit of its relative frequency in a large number of trials. This interpretation needs the possibility that a given game or experiment can be repeated (with the same physical conditions) several times. In this book, we take this point of view. We assume that there exist reproducible experiments, which, under very similar initial conditions, produce different events (denoted by E_i, a Boolean variable that may be true or false, when the event i is realized or not, respectively). Within the frequentist definition, the probability assigned to any event E_i is given by the ratio between the number of events n_i in which E_i happened and the total number of trials N, when N becomes very large:

$$P(E_i \text{ is true}) \equiv P(i) = \lim_{N\to\infty} \frac{n_i}{N}; \tag{2.1}$$

clearly, the probability is a non-negative number:

$$P(i) \geq 0. \tag{2.2}$$

In the following, we give a brief overview of the theory of probability. A clear and comprehensive treatment of this subject is given in the book by Gnedenko (2014), which also includes the assiomatic approach developed by the Russian mathematician Andrey Nikolaevich Kolmogorov in 1933.

2.2 Events and Probability

Here, we describe some simple properties of events. Two events E_i and E_j are said to be *mutually exclusive* if and only if the occurrence of E_i implies that E_j does not occur and vice versa. If E_i and E_j are mutually exclusive, we have that:

$$P(E_i \text{ is true and } E_j \text{ is true}) = 0, \tag{2.3}$$

$$P(E_i \text{ is true or } E_j \text{ is true}) = P(i) + P(j). \tag{2.4}$$

The different outcomes of an experiment represent mutually exclusive events (for example, when tossing a coin, head and tail represent two mutually exclusive

https://doi.org/10.1017/9781316417041.003 Published online by Cambridge University Press

events). Clearly, if the number of all possible mutually exclusive events is M, then the sum of their probability over the whole space of outcomes is given by:

$$\sum_{i=1}^{M} P(i) = 1. \tag{2.5}$$

Once for a given experiment all the M mutually exclusive events are classified, each realization of the experiment is specified by a single integer i, such that E_i is verified. Therefore, we can define a *random variable* X_i, as a real-valued function associated to any possible successful event E_i. The simplest random variable is the *characteristic* random variable $X_i^{[j]}$:

$$X_i^{[j]} = \begin{cases} 1 & \text{if } i = j \ (E_j \text{ is true}) \\ 0 & \text{if } i \neq j \ (E_j \text{ is false}) \end{cases} \tag{2.6}$$

in other words, the characteristic random variable $X_i^{[j]}$ is non-zero only if the event E_j is successful: for example, if we bet that the number 36 will show up in the roulette game, the successful event, associated to a winning bet, is the appearance of 36, while all the other numbers would give rise to a loosing bet. A different example of a random variable is given by the actual outcome after rolling a die, i.e., the number that shows up:

$$X_i = i \text{ if } E_i \text{ is true}, \tag{2.7}$$

for $i = 1, \ldots, 6$.

Let us now discuss the case of composite events. For example, rolling two dice is an experiment that can be characterized by $E_i^{(1)}$ and $E_j^{(2)}$, where $E_j^{(1)}$ ($E_j^{(2)}$) refers to the possible outcomes of the first (second) die. For composite events, the probability is labeled by more than one index, in particular the *joint probability* of two events $P(i,j)$ is defined as:

$$P_{\text{joint}}(i,j) = P(E_i^{(1)} \text{ is true and } E_j^{(2)} \text{ is true}). \tag{2.8}$$

In order to obtain the probability for one variable alone (say $E_i^{(1)}$), consistently with the frequentist definition (2.1), we have to sum the joint probability over all values of the other variable (say $E_j^{(2)}$). In this way, we get the *marginal probability*:

$$P_1(i) = \sum_{j=1}^{M} P_{\text{joint}}(i,j), \tag{2.9}$$

which represents the probability of obtaining the variable $E_i^{(1)}$, without caring about the outcome of the variable $E_j^{(2)}$. Instead, we can be interested into the probability of obtaining $E_i^{(1)}$, once we know the outcome of the variable $E_j^{(2)}$; this is the *conditional probability* $\omega(i|j)$:

https://doi.org/10.1017/9781316417041.003 Published online by Cambridge University Press

$$\omega(i|j) = \frac{P_{\text{joint}}(i,j)}{\sum_{i=1}^{M} P_{\text{joint}}(i,j)}, \tag{2.10}$$

which is normalized:

$$\sum_{i=1}^{M} \omega(i|j) = 1. \tag{2.11}$$

Since $\sum_{i=1}^{M} P_{\text{joint}}(i,j) = P_2(j)$, we have that:

$$\omega(i|j) = \frac{P_{\text{joint}}(i,j)}{P_2(j)}, \tag{2.12}$$

and, therefore, the joint probability can be expressed as the product of the marginal probability $P_2(j)$ times the conditional probability $\omega(i|j)$:

$$P_{\text{joint}}(i,j) = \omega(i|j)P_2(j). \tag{2.13}$$

Similarly, we can also obtain that:

$$P_{\text{joint}}(i,j) = \omega(j|i)P_1(i). \tag{2.14}$$

By combining Eqs. (2.13) and (2.14), we obtain the Bayes formula:

$$\omega(j|i)P_1(i) = \omega(i|j)P_2(j). \tag{2.15}$$

One random variable i is *independent* from the other one j whenever the conditional probability $\omega(i|j)$ does not depend on j; in this case, we have that $P_1(i) = \sum_{j=1}^{M} \omega(i|j)P_2(j) = \omega(i|j)$, which inserted into Eq. (2.13), gives:

$$P_{\text{joint}}(i,j) = P_1(i)P_2(j); \tag{2.16}$$

notice that, from Eq. (2.14), we also have that $P_2(j) = \omega(j|i)$, implying that also the variable j is independent from i.

The generalization of the above definitions holds obviously also for random variables x defined on the continuum, namely variables that assume any real value within a given interval and not only discrete values. In this case, the probability to obtain a particular value x is generically vanishing, while there is a finite probability to find the random variable within a finite range (a, b). In particular, we can consider the probability that x is smaller than y, where y is a given fixed real number:

$$P(x \le y) = F(y) = \int_{-\infty}^{y} dx\, \mathcal{P}(x), \tag{2.17}$$

which is called the *cumulative probability* of the random variable x. Clearly $F(+\infty) = 1$. The *probability density* or *distribution function* $\mathcal{P}(x)$ is then given by:

$$\mathcal{P}(x) = \left.\frac{dF(y)}{dy}\right|_{y=x}, \tag{2.18}$$

https://doi.org/10.1017/9781316417041.003 Published online by Cambridge University Press

which satisfies the following properties:

$$\mathcal{P}(x) \geq 0, \tag{2.19}$$

$$\int_{-\infty}^{+\infty} dx\, \mathcal{P}(x) = 1. \tag{2.20}$$

Notice that the above definitions can be also used in the case of discrete random variables, by taking:

$$\mathcal{P}(x) = \sum_{i=1}^{M} P(i)\delta(x - X_i). \tag{2.21}$$

Therefore, in the following, we will consider the formalism of continuous random variables, having in mind that the discrete case can be simply obtained by considering Eq. (2.21).

As before, the marginal probability of the variable x is defined as:

$$\mathcal{P}_1(x) = \int_{-\infty}^{+\infty} dy\, \mathcal{P}_{\text{joint}}(x, y), \tag{2.22}$$

and the conditional probability of obtaining x once y is known is given by:

$$\omega(x|y) = \frac{\mathcal{P}_{\text{joint}}(x, y)}{\int_{-\infty}^{+\infty} dx\, \mathcal{P}_{\text{joint}}(x, y)}, \tag{2.23}$$

such that:

$$\mathcal{P}_{\text{joint}}(x, y) = \omega(x|y)\mathcal{P}_2(y) = \omega(y|x)\mathcal{P}_1(x). \tag{2.24}$$

2.3 Moments of the Distribution: Mean Value and Variance

For any random variable x, we can define its *mean value* or *expected value*:

$$\mu \equiv \langle x \rangle = \int_{-\infty}^{+\infty} dx\, x\, \mathcal{P}(x). \tag{2.25}$$

Within the frequentist approach, the expected value of a random variable is just the average value of several repetitions of the same experiment. For example, when rolling a die, the expected value is 3.5. For the characteristic random variable $X_i^{[j]}$ of the event E_j, we simply have $\langle X_i^{[j]} \rangle = P(j)$. More generally, the n-th *moment* of the distribution is defined as the expected value of the n-th power of x:

$$\langle x^n \rangle = \int_{-\infty}^{+\infty} dx\, x^n\, \mathcal{P}(x). \tag{2.26}$$

https://doi.org/10.1017/9781316417041.003 Published online by Cambridge University Press

The first moment $\langle x\rangle$ is equal to the mean value μ. The second moment allows us to define a particularly important quantity, which is the *variance*:

$$\sigma^2 \equiv \langle x^2\rangle - \langle x\rangle^2 = \int_{-\infty}^{+\infty} dx\ (x-\langle x\rangle)^2\ \mathcal{P}(x). \tag{2.27}$$

The variance is a non-negative quantity that can be zero only when all the events having a non-vanishing probability give the same value for the variable x. In other words, whenever the variance is zero, the random character of the variable is completely lost and the experiment becomes perfectly predictable. The square root of the variance is a measure of the dispersion of the random variable and is called *standard deviation*. Notice that the existence of the variance, and of higher moments as well, is not *a priori* guaranteed in the continuous case. Indeed, the probability density $\mathcal{P}(x)$ has to decrease sufficiently fast for $x \to \pm\infty$ in order for the corresponding integrals to exist.

In the case we have two random variables, we can define their *covariance* by:

$$\sigma_{xy}^2 \equiv \langle xy\rangle - \langle x\rangle\langle y\rangle = \int_{-\infty}^{+\infty} dx \int_{-\infty}^{+\infty} dy\ (x-\langle x\rangle)(y-\langle y\rangle)\ \mathcal{P}_{\text{joint}}(x,y); \tag{2.28}$$

here, we must notice that the quantity σ_{xy}^2 is not guaranteed to be positive; nevertheless, in analogy to the variance, we prefer to keep the notation with the square. Obviously, for two independent variables, we have that $\sigma_{xy}^2 = 0$, as obtained from the fact that $\mathcal{P}_{\text{joint}}(x,y) = \mathcal{P}_x(x)\mathcal{P}_y(y)$.

Finally, we can also consider expected values of another random variable $y = f(x)$:

$$\langle f(x)\rangle = \int_{-\infty}^{+\infty} dx\, f(x)\ \mathcal{P}(x). \tag{2.29}$$

An important quantity related to the probability density $\mathcal{P}(x)$ is the *characteristic function* $\phi_x(t)$, which is the expected value of e^{ixt}, or equivalently the Fourier transform of the probability density:

$$\phi_x(t) = \langle e^{ixt}\rangle = \int_{-\infty}^{+\infty} dx\ e^{ixt}\ \mathcal{P}(x). \tag{2.30}$$

For small t, if the second moment $\langle x^2\rangle$ is finite, one can expand the exponential up to second order in t, obtaining:

$$\phi_x(t) = 1 + i\langle x\rangle t - \frac{\langle x^2\rangle}{2}t^2 + o(t^2), \tag{2.31}$$

where $o(t^2)$ are terms that are smaller than t^2: if the third moment is also finite, the next term will be proportional to t^3, while if $\langle x^3\rangle = \infty$, this term will have an

https://doi.org/10.1017/9781316417041.003 Published online by Cambridge University Press

intermediate power between 2 and 3. In turn, the probability density $\mathcal{P}(x)$ is the Fourier transform of the characteristic function:

$$\mathcal{P}(x) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} dt\, e^{-ixt}\, \phi_x(t). \tag{2.32}$$

Here, we would like to report few important examples of distribution functions. For a discrete random variable, the simplest possible case is given by the Bernoulli distribution, named after the Swiss scientist Jacob Bernoulli. In this case, the random variable takes the value 1 with "success" probability $\mathcal{P}$ and the value 0 with "failure" probability $\mathcal{Q} = 1 - \mathcal{P}$. It can be used to represent a coin toss where 1 and 0 would represent head and tail, respectively. In particular, fair coins have $\mathcal{P} = \mathcal{Q} = 1/2$. The mean of the Bernoulli distribution is equal to $\mathcal{P}$, while the variance is $\mathcal{P}\mathcal{Q}$. Its characteristic function is given by $\phi_{\text{bernulli}}(t) = \mathcal{Q} + \mathcal{P}e^{it}$.

Another example is given by the binomial distribution, which describes the number of successes in a sequence of N independent trials (experiments), each of which yields success with probability $\mathcal{P}$ and failure with probability $\mathcal{Q}$ (e.g., N trials of the Bernoulli variable). The probability of obtaining k successes in N trials is given by:

$$P_{\text{binomial}}(k) = \frac{N!}{k!\ (N-k)!} \mathcal{P}^k \mathcal{Q}^{N-k}. \tag{2.33}$$

The mean of the binomial distribution is $N\mathcal{P}$, while the variance is $N\mathcal{P}\mathcal{Q}$. The characteristic function is $\phi_{\text{binomial}}(t) = [\phi_{\text{bernulli}}(t)]^N$. In the limit of $N \to \infty$ and $\mathcal{P} \to 0$ with $N\mathcal{P} = \lambda$, $P_{\text{binomial}}(k)$ approaches the Poisson distribution:

$$P_{\text{poisson}}(k) = e^{-\lambda} \frac{\lambda^k}{k!}, \tag{2.34}$$

which describes the probability of an unlike event ($\mathcal{P} \to 0$) in a large number of trials ($N \to \infty$). Both the mean and the variance of the Poisson distribution are equal to λ. Its characteristic function is $\phi_{\text{poisson}}(t) = \exp[\lambda(e^{it} - 1)]$.

For continuous random variables, the simplest possible distribution function is the uniform one, in which all the values of x in the interval $[a, b]$ are equiprobable:

$$\mathcal{P}_{\text{uniform}}(x) = \begin{cases} \frac{1}{(b-a)} & \text{if } a \le x \le b, \\ 0 & \text{otherwise.} \end{cases} \tag{2.35}$$

The mean of the uniform distribution is equal to $(b + a)/2$, while the variance is $(b - a)^2/12$. Its characteristic function is $\phi_{\text{uniform}}(t) = \left(e^{bt} - e^{at}\right)/[t(b-a)]$.

Finally, a pivotal role in the probability theory is given by the Gaussian (or normal) distribution, named after the German scientist Carl Friederich Gauss:

$$\mathcal{P}_{\text{gauss}}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{2.36}$$

https://doi.org/10.1017/9781316417041.003 Published online by Cambridge University Press

where, μ and σ^2 denote the mean and the variance, respectively. Its characteristic function, which will be used in the following, is given by:

$$\phi_{\text{gauss}}(t) = e^{i\mu t} e^{-\frac{t^2}{2}\sigma^2}. \tag{2.37}$$

2.4 Changing Random Variables

In this section, we briefly show how the probability density is modified when we apply a generic transformation to a random variable. Suppose that we have a random variable x and we construct another random variable y that is a function of x:

$$y = f(x), \tag{2.38}$$

where, for simplicity, we consider a monotonic function $f(x)$, with positive derivative, i.e., $df(x)/dx > 0$, such that $f(x_2) = y_2 > f(x_1) = y_1$ for $x_2 > x_1$ (the case with $df(x)/dx < 0$ can be treated similarly); in the most general case of a non-monotonic function, we can take separately all the small intervals where the function is monotonic. Then, the probability to find x in a given interval $x_1 < x < x_2$ must be equal to the probability to find y in the interval $y_1 < y < y_2$:

$$P_y(y_1 < y < y_2) = P_x(x_1 < x < x_2). \tag{2.39}$$

Then, for a small interval $dx = (x_2 - x_1)$ for which $dy = df(x)$, we have that:

$$\mathcal{P}_y(y)dy = \mathcal{P}_x(x)dx. \tag{2.40}$$

This equation gives a punctual relation between the values of the probability densities at x and y, we can integrate both sides and get:

$$\int_{-\infty}^{y} ds\, \mathcal{P}_y(s) = \int_{-\infty}^{x} dt\, \mathcal{P}_x(t). \tag{2.41}$$

The usefulness of this relation is that it allows one to obtain the probability density for the variable y, i.e., $\mathcal{P}_y(y)$, once we know the probability density of the variable x, i.e., $\mathcal{P}_x(x)$, and the relation between x and y, i.e., the function $f(x)$. Notice that, whenever $df(x)/dx < 0$, the only modification comes from exchanging the limits of integration in the r.h.s. of Eq. (2.41).

The simplest application of the above relations is obtained when considering a linear transformation:

$$y = a + bx; \tag{2.42}$$

then, we simply get:

$$\mathcal{P}_y(y) = \frac{1}{b}\mathcal{P}_x\left(\frac{y-a}{b}\right). \tag{2.43}$$

https://doi.org/10.1017/9781316417041.003 Published online by Cambridge University Press

Therefore, we immediately obtain that, if the random variable x has mean μ_x and variance σ_x^2, then the transformed random variable y has $\mu_y = a + b\mu_x$ and $\sigma_y^2 = b^2\sigma_x^2$.

2.5 The Chebyshev's Inequality

The Chebyshev's inequality gives a simple bound for the probability to obtain a result x which lies far from the mean value $\langle x \rangle$; the precise statement is that no more than $1/k^2$ of the distribution's values can be more than k standard deviations away from the mean value:

$$P\left(|x - \langle x \rangle| > k\sigma\right) \leq \frac{1}{k^2}, \tag{2.44}$$

or equivalently, by taking $\epsilon = k\sigma$:

$$P\left(|x - \langle x \rangle| > \epsilon\right) \leq \frac{\sigma^2}{\epsilon^2}. \tag{2.45}$$

The proof of the Chebyshev's inequality is very simple. Indeed, by definition:

$$P\left(|x - \langle x \rangle| > \epsilon\right) = \int_{-\infty}^{\langle x \rangle - \epsilon} dx\, \mathcal{P}(x) + \int_{\langle x \rangle + \epsilon}^{+\infty} dx\, \mathcal{P}(x); \tag{2.46}$$

in both these two intervals $[(x - \langle x \rangle)/\epsilon]^2 \geq 1$, thus we have:

$$\begin{aligned} P\left(|x - \langle x \rangle| > \epsilon\right) &\leq \int_{-\infty}^{\langle x \rangle - \epsilon} dx \left(\frac{x - \langle x \rangle}{\epsilon}\right)^2 \mathcal{P}(x) + \int_{\langle x \rangle + \epsilon}^{+\infty} dx \left(\frac{x - \langle x \rangle}{\epsilon}\right)^2 \mathcal{P}(x) \\ &\leq \int_{-\infty}^{+\infty} dx \left(\frac{x - \langle x \rangle}{\epsilon}\right)^2 \mathcal{P}(x) = \frac{\sigma^2}{\epsilon^2}. \end{aligned} \tag{2.47}$$

The Chebyshev's bound is very general and does not assume any form of the probability distribution $\mathcal{P}(x)$; for this reason, it gives a rather weak bound on the probability to find a large fluctuation of x, i.e., the probability to find an event very far from its expected value. For specific distributions, it is possible to obtain much stronger bounds. For example, in Table 2.1, we report a comparison between the Chebyshev's bound of Eq. (2.44) and the actual values that are valid for the Gaussian distribution of Eq. (2.36).

2.6 Summing Independent Random Variables

Here, we treat the case of several independent random variables $x_1, \ldots, x_N$ for which, generalizing the arguments of section 2.2, the joint probability is the product of the marginal ones, i.e., $\mathcal{P}(x_1, \ldots, x_N) = \mathcal{P}_1(x_1) \ldots \mathcal{P}_N(x_N)$. Then, it is often

https://doi.org/10.1017/9781316417041.003 Published online by Cambridge University Press

Table 2.1. *Probability that a random variable has a fluctuation away from its mean value larger than k standard deviations σ. The results of the Chebyshev's bound of Eq. (2.44) (left column) are compared with the actual values obtained from the Gaussian distribution of Eq. (2.36) (right column).*

k	Chebyshev's bound	Gaussian value
1	1.0000	0.3173
2	0.2500	0.0455
3	0.1111	0.0027
4	0.0625	6.3×10^{-5}
5	0.0400	5.8×10^{-7}
6	0.0278	2.0×10^{-9}
7	0.0204	2.5×10^{-12}
8	0.0156	1.2×10^{-15}

useful to consider a "new" random variable that is the sum of them. The main motivation of summing (independent and identically distributed) random variables is due to the fact that important and useful properties about the distribution function of the sum can be obtained in the limit of large N, i.e., for $N \to \infty$ (Gnedenko and Kolmogorov, 1954). Let us start by taking z that is the sum of two random variables:

$$z = x_1 + x_2. \tag{2.48}$$

Its probability density is simply given by:

$$\mathcal{P}_z(z) = \int_{-\infty}^{+\infty} dx_1 \int_{-\infty}^{+\infty} dx_2 \; \mathcal{P}_1(x_1)\, \mathcal{P}_2(x_2)\, \delta(z - x_1 - x_2), \tag{2.49}$$

where the delta-function enforces the fact that z is the sum of x_1 and x_2. Therefore, by performing the integral over x_2, we have that $\mathcal{P}_z(z)$ is simply given by the convolution of the probability densities:

$$\mathcal{P}_z(z) = \int_{-\infty}^{+\infty} dx_1 \; \mathcal{P}_1(x_1)\, \mathcal{P}_2(z - x_1), \tag{2.50}$$

whose meaning is the following: the probability for a given value z is obtained by summing (or integrating) over all the possible outcomes of the variable x_1 the probability of finding x_1 times the probability that $x_2 = z - x_1$. From Eq. (2.50), we have that the characteristic function of the sum of two independent random variables x_1 and x_2 is the product of the two characteristic functions:

$$\phi_z(t) = \phi_1(t)\phi_2(t). \tag{2.51}$$

https://doi.org/10.1017/9781316417041.003 Published online by Cambridge University Press

Let us see some examples in which we can easily obtain the probability of a sum of random variables. The simplest case is given when two random variables are uniformly distributed in [0, 1], for which we get the "triangular" distribution:

$$\mathcal{P}_z(z) = \begin{cases} 0 & \text{if } z \leq 0, \\ z & \text{if } 0 \leq z \leq 1, \\ 2 - z & \text{if } 1 \leq z \leq 2, \\ 0 & \text{if } z \geq 2. \end{cases} \tag{2.52}$$

As expected, it is impossible to get a sum that is negative or exceeds 2, while the most probable value (which also coincides with the mean value) is $z = 1$. This example is the continuous version of the case where we roll two dice and consider the sum of the two outcomes: the most probable sum is 7, while it is impossible to get something that is smaller than 2 or larger than 12.

Then, the sum of three uniformly distributed random variables $z = x_1 + x_2 + x_3$ can be easily obtained by adding together two of them and get the "triangular" distribution and then adding the third one using Eq. (2.50). This procedure gives the following distribution for the sum of three uniformly distributed variables:

$$\mathcal{P}_z(z) = \begin{cases} 0 & \text{if } z \leq 0, \\ \frac{z^2}{2} & \text{if } 0 \leq z \leq 1, \\ \frac{3}{4} - \left(z - \frac{3}{2}\right)^2 & \text{if } 1 \leq z \leq 2, \\ \frac{(3-z)^2}{2} & \text{if } 2 \leq z \leq 3, \\ 0 & \text{if } z \geq 3. \end{cases} \tag{2.53}$$

Here, the most probable value is $z = 3/2$. The trend of this procedure is clear: when adding up more and more random variables, both the mean value and variance of the distribution increase. In general, the probability of the sum of random variables depends upon the number of the variables. One remarkable exception is given by Gaussian variables, for which the sum of two random numbers is still Gaussian. Indeed, by using Eq. (2.51) and the form of the characteristic function of the Gaussian distribution (2.37), we have that (indicating by μ_1 and μ_2 the means and by σ_1^2 and σ_2^2 the variances of the two variables):

$$\phi_z(t) = e^{i(\mu_1+\mu_2)t} e^{-\frac{t^2}{2}(\sigma_1^2+\sigma_2^2)}, \tag{2.54}$$

which is just the characteristic function of a Gaussian variable with mean $\mu_z = \mu_1 + \mu_1$ and variance $\sigma_z^2 = \sigma_1^2 + \sigma_2^2$.

If the random variables have generic distribution functions, then the distribution function of the sum of them will have a very complicated form. The remarkable fact is that, under very general conditions, it is possible to obtain an asymptotically *exact* form for $\mathcal{P}_z(z)$ when the number N of independent variables becomes very

https://doi.org/10.1017/9781316417041.003 Published online by Cambridge University Press

large. Before demonstrating this fundamental aspect (which goes under the name of central limit theorem), we would like to discuss some relevant issues about the mean and variance of the sum of random variables. Given the fact that both the mean and the variance are proportional to N, it is often useful to consider the average of independent and equally distributed random variables:

$$\overline{x} = \frac{1}{N} \sum_i x_i. \tag{2.55}$$

Even though, in general cases, it is not easy to work out the explicit form of the probability distribution of $\overline{x}$, both the mean and the variance of $\overline{x}$ can be easily computed. Indeed, all the N terms in the sum give an identical contribution, equal to $\langle x \rangle$, thus giving:

$$\langle \overline{x} \rangle = \langle x \rangle, \tag{2.56}$$

namely, the mean value of the average $\overline{x}$ coincides with the mean value of the single trial (experiment). We would like to emphasize that Eq. (2.56) holds also in the case where the random variables x_i are not independent. In order to compute the variance of $\overline{x}$, we simply notice that, by taking the expected value of $\overline{x}^2$, we get:

$$\begin{aligned} \langle \overline{x}^2 \rangle &= \frac{1}{N^2} \sum_{i,j} \langle x_i x_j \rangle = \frac{1}{N^2} \left[\sum_i \langle x_i^2 \rangle + \sum_{i \neq j} \langle x_i \rangle \langle x_j \rangle \right] \\ &= \frac{1}{N^2} \left[N \langle x^2 \rangle + N(N-1) \langle x \rangle^2 \right], \end{aligned} \tag{2.57}$$

where we have used the fact that the variables are independent, leading to $\langle x_i x_j \rangle = \langle x_i \rangle \langle x_j \rangle$ for $i \neq j$; then, $\langle \overline{x}^2 \rangle$ has two contributions, the first one, coming from the terms with $i = j$, gives N terms that do not depend upon i (i.e., $\langle x_i^2 \rangle = \langle x^2 \rangle$); the second one, coming from the terms with $i \neq j$, gives $N(N-1)$ terms that are all equal (i.e., $\langle x_i \rangle \langle x_j \rangle = \langle x \rangle^2$). Therefore, the variance of $\overline{x}$ is given by:

$$\sigma_{\overline{x}}^2 = \langle \overline{x}^2 \rangle - \langle \overline{x} \rangle^2 = \frac{\sigma^2}{N}, \tag{2.58}$$

where $\sigma^2 = \langle x^2 \rangle - \langle x \rangle^2$ is the variance of the single random variable x. Therefore, for large N, the random variable $\overline{x}$, corresponding to averaging a large number of realizations of the same experiment, will have a very narrow distribution function, since $\sigma_{\overline{x}}^2 \to 0$ for $N \to \infty$, with the same mean value of the original random variables. In other words, almost all possible average measurements (each of them done by N different realizations of the same experiment) give a value for $\overline{x}$ that is closer to the true mean value than the single experiment. This important fact can be obtained in a more rigorous way by applying the Chebyshev's inequality of Eq. (2.45) to the variable $\overline{x}$:

https://doi.org/10.1017/9781316417041.003 Published online by Cambridge University Press

$$P\left(\left|\frac{1}{N}\sum_i x_i - \langle x\rangle\right| > \epsilon\right) \leq \frac{\sigma_{\overline{x}}^2}{\epsilon^2} = \frac{\sigma^2}{N\epsilon^2}, \tag{2.59}$$

which directly gives:

$$\lim_{N\to\infty} P\left(\left|\frac{1}{N}\sum_i x_i - \langle x\rangle\right| > \epsilon\right) = 0. \tag{2.60}$$

This result is nothing but the so-called *weak law of large numbers*, which essentially states that for any non-zero margin ϵ, for a sufficiently large N, there will be a very high probability that the average $\overline{x}$ will be close enough to the expected value, i.e., within the margin ϵ.

Eqs. (2.56) and (2.58) show that the average of many independent random variables gives an unbiased and very accurate estimation of the true expectation value. Since the exact variance of the distribution σ^2 is not generally known, it is useful to define an estimator for it, which allows us to estimate $\sigma_{\overline{x}}^2$ of Eq. (2.58). For that, we can consider:

$$s^2 = \frac{1}{N}\sum_i (x_i - \overline{x})^2 = \frac{1}{N}\sum_i x_i^2 - \left(\frac{1}{N}\sum_i x_i\right)^2, \tag{2.61}$$

which is also a random variable with:

$$\langle s^2\rangle = \frac{1}{N}\sum_i \langle x_i^2\rangle - \frac{1}{N^2}\sum_{i,j}\langle x_i x_j\rangle, \tag{2.62}$$

whenever the variables x_i are independent and identically distributed, we obtain:

$$\langle s^2\rangle = \left(1 - \frac{1}{N}\right)\left(\langle x^2\rangle - \langle x\rangle^2\right) = \left(\frac{N-1}{N}\right)\sigma^2. \tag{2.63}$$

Therefore, for $N \to \infty$, s^2 gives an asymptotically exact estimation of the variance. Notice that a slightly better estimation of the variance for any finite values of N should be given by taking $N/(N-1)s^2$ instead of s^2.

Finally, suppose that the random variable x_i is just the characteristic random variable of a given event E_j, namely $x_i \equiv X_i^{[j]}$. For this random variable, we have already noticed that the mean value is the probability of the event E_j, namely $\langle X_i^{[j]}\rangle = P(j)$. In addition, in view of the previous discussion, the mean of the random variable $\overline{x}$, obtained by averaging N independent realizations of the same experiment, gives an estimate of $P(j)$, with a standard deviation that decreases with $1/\sqrt{N}$, see Eq. (2.58). This uncertainty can be made arbitrarily small, by increasing N, so that the probability $P(j)$ of the event E_j is a well defined quantity in the limit of $N \to \infty$. This fact consistently justifies the definition of Eq. (2.1), which is the

https://doi.org/10.1017/9781316417041.003 Published online by Cambridge University Press

basis of the frequentist approach to probability. Notice that, within this scheme, the concept of probability is related to the *reproducibility* of the experiments.

2.7 The Central Limit Theorem

Let us finally prove a very important theorem in the theory of probability, the so-called *central limit theorem*, which provides the asymptotic probability distribution of the sum over a large number of random variables x_i, which are independent and equally distributed with probability $\mathcal{P}(x)$. The importance of this theorem relies on the fact that the asymptotic form of the distribution function is given under very general conditions, i.e., regardless of the specific form of $\mathcal{P}(x)$. The only requirement is that the variance is finite. As a first step, we define $y_i = x_i - \langle x \rangle$, which are still independent and equally distributed random variables, having $\langle y_i \rangle = 0$ by definition. Then, we construct the random variable Y:

$$Y = \frac{1}{\sqrt{N}} \sum_i y_i = \sqrt{N}(\bar{x} - \langle x \rangle), \tag{2.64}$$

which also has a vanishing expected value:

$$\langle Y \rangle = 0. \tag{2.65}$$

The characteristic function of Y is given by:

$$\phi_Y(t) = \left\langle \exp\left(\frac{it}{\sqrt{N}} \sum_i y_i\right)\right\rangle = \left\langle \prod_i \exp\left(\frac{it}{\sqrt{N}} y_i\right)\right\rangle = \left[\phi_y\left(\frac{t}{\sqrt{N}}\right)\right]^N, \tag{2.66}$$

where we have used the fact that all y_i are independent and equally distributed, so that all terms give the same factor. For small values of t, we can expand $\phi_y(\frac{t}{\sqrt{N}})$ up to second order by using Eq. (2.31) with $\mu = 0$:

$$\phi_Y(t) = \left[1 - \frac{\sigma^2 t^2}{2N} + o\left(\frac{t^2}{N}\right)\right]^N. \tag{2.67}$$

When considering the limit of a very large number of random variables, we get:

$$\lim_{N\to\infty} \phi_Y(t) = \exp\left(-\frac{\sigma^2}{2} t^2\right), \tag{2.68}$$

which is the characteristic function of a Gaussian random variable with mean $\mu = 0$ and variance σ^2:

$$\mathcal{P}(Y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{Y^2}{2\sigma^2}\right). \tag{2.69}$$

https://doi.org/10.1017/9781316417041.003 Published online by Cambridge University Press

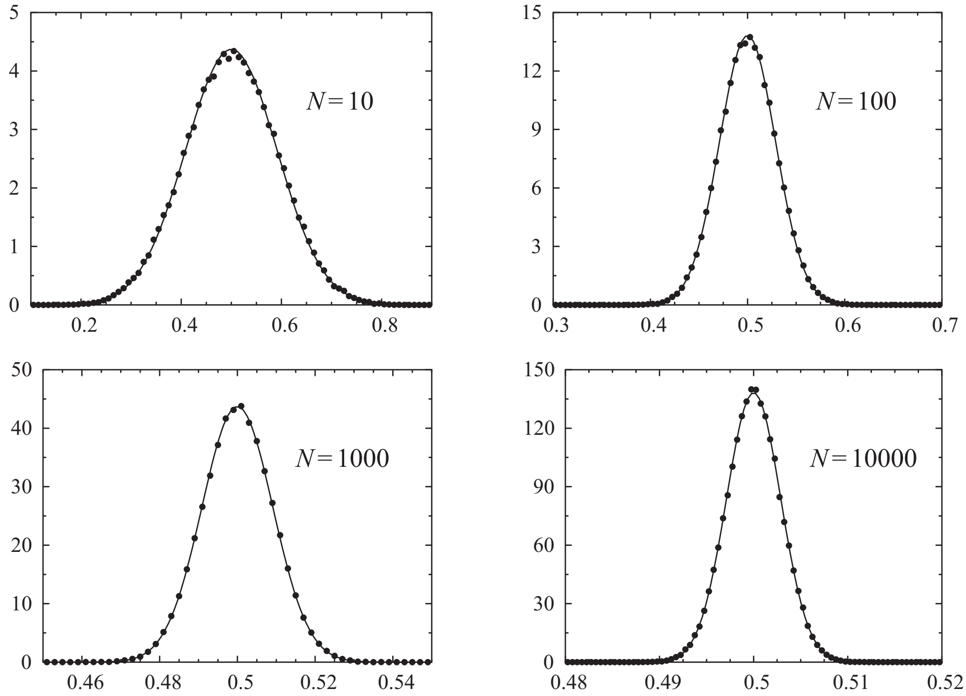

Figure 2.1 The probability distribution of $\overline{x} = 1/N \sum_i x_i$, generated over 10^5 realizations of $\overline{x}$, for different values of N. The expected Gaussian distributions, with $\mu = 0.5$ and $\sigma^2 = 1/(12N)$ are also reported.

By using the results of section 2.4, we have that also $\overline{x} = \langle x \rangle + Y/\sqrt{N}$ is a Gaussian variable with mean equal to $\langle x \rangle$ and variance equal to σ^2/N:

$$\mathcal{P}(\overline{x}) = \sqrt{\frac{N}{2\pi\sigma^2}} \exp\left[-\frac{N(\overline{x} - \langle x \rangle)^2}{2\sigma^2}\right]. \tag{2.70}$$

In Fig. 2.1, we report few examples in which we generated 10^5 averages $\overline{x}$ of Eq. (2.55), obtained by summing N random variables that are uniformly distributed in $[0, 1)$. In particular, we divide the interval $[0, 1)$ into L small sub-intervals of width $1/L$ and compute the number of times that $\overline{x}$ falls in a given interval (normalized to the total number of trials). This quantity approaches very rapidly the Gaussian distribution (with the expected mean and variance), as predicted by the central limit theorem: already for $N = 10$, the distribution is essentially indistinguishable from a Gaussian.

The importance of the central limit theorem lies in the fact that the details (e.g., all the moments with $n > 2$) of the original distribution function $\mathcal{P}(x)$ do not contribute to the form of the asymptotic distribution function of the average $\overline{x}$, which is then universal. Most importantly, summing many random variables gives

https://doi.org/10.1017/9781316417041.003 Published online by Cambridge University Press

rise to a random variable that has the same mean value of the original ones but has a much smaller variance, which tends to zero when $N \to \infty$. In other words, the random variable $\bar{x}$ becomes less and less fluctuating when increasing N, eventually becoming a deterministic number for $N \to \infty$. Therefore, the estimation of the mean is by far much more precise when summing many random variables than considering the single one. As we will discuss in the next chapter, the central limit theorem represents the heart of any Monte Carlo method to evaluate integrals.

We emphasize the fact that the Gaussian form of Eq. (2.70) should be taken with a caveat. Indeed, Eq. (2.70) holds only in the neighborhood of its maximum, i.e., for $(\bar{x} - \langle x \rangle) = O(\sigma/\sqrt{N})$, and not in the tails of the distribution, where *large deviations* usually appear. An extension of the central limit theorem is given by the so-called *large deviations theory*, which gives information not only to small deviations, i.e., $O(1/\sqrt{N})$, but also to rare events that are far away from the typical values.

We would like to finish this part by considering an application of the central limit theorem. Instead of taking the average of N random variables, let us consider the product of them:

$$z = \prod_{i=1}^{N} x_i. \tag{2.71}$$

Both the average and the variance of z can be worked out in the limit of large N. Indeed:

$$\xi = \ln z = \sum_{i=1}^{N} \ln x_i \tag{2.72}$$

is the sum of N random variables and, therefore, has a Gaussian distribution for $N \to \infty$:

$$\mathcal{P}(\xi) = \sqrt{\frac{1}{2\pi N\sigma^2}} \exp\left[-\frac{(\xi - \langle \xi \rangle)^2}{2N\sigma^2}\right], \tag{2.73}$$

where $\langle \xi \rangle = N\langle \ln x \rangle$ and $\sigma^2 = \langle (\ln x)^2 \rangle - \langle \ln x \rangle^2$. Thus, we obtain that:

$$\langle z \rangle = \langle e^{\xi} \rangle \propto \exp\left(\langle \xi \rangle + \frac{N}{2}\sigma^2\right), \tag{2.74}$$

$$\langle z^2 \rangle = \langle e^{2\xi} \rangle \propto \exp\left(2\langle \xi \rangle + 2N\sigma^2\right), \tag{2.75}$$

which imply that for $N \to \infty$:

$$\frac{\sqrt{\langle z^2 \rangle - \langle z \rangle^2}}{\langle z \rangle} \approx \exp\left(\frac{N}{2}\sigma^2\right) \to \infty. \tag{2.76}$$

Therefore, in contrast to the sum of random variables, the product of them has a variance which diverges much faster than its expectation value.

https://doi.org/10.1017/9781316417041.003 Published online by Cambridge University Press