1 Introduction

Many important everyday phenomena in nature appear to be unpredictable or random. Fluctuations in the neutral winds in the atmosphere are representative examples of large-scale phenomena. Examples on small scales are fluctuations in electric circuits, and random movements of tiny grains of pollen suspended in liquid; so-called Brownian motions (MacDonald, 1962). This and a number of related problems and phenomena will be discussed in some detail in the following chapters.

Although the concept of randomness may seem intuitively clear, its actual definition is somewhat ambiguous. Randomness is often associated with unpredictability, but the fact that one observer is unable to predict or comprehend a certain sequence of signals or events does not preclude the possibility that it seems perfectly transparent to another who has more *a priori* information available.[1] As an illustration of this point, consider for instance the sequence of numbers

111111111111111, 10000, 121, 100, 31, 24, 22, ○, 17, 16, 15, 14, 13, 12, 11, 10, . . .

and predict the number at the position indicated by ○. Even though the sequence has been ordered in some sense and does not appear random, it does not, on the other hand, seem to be regular in any way (in particular not when we are told that the next symbol in the sequence is G, and so are actually also all the following ones!). However, with the proper *a priori* information the entire sequence is perfectly meaningful, and actually quite simple to comprehend.

A time-varying signal is deterministic, i.e. completely predictable, provided that it is infinitely many times differentiable and known *a priori* in a small time interval. Then, by a Taylor expansion, it can in principle be described with arbitrary accuracy to arbitrary later times. Unpredictability is therefore here related to discontinuities either in the functional values or in time derivatives at some order invalidating this expansion. Quite formally this is so, and it agrees also with intuitive expectations of random functions looking ragged. In reality it is not possible to determine all derivatives with the desired accuracy on the basis of a given time sequence and the ideal prediction outlined here is not feasible.

For a wide class of physical systems we have to accept that information can be available only on a certain level, even in the classical limit. Even though we assume that the forces acting on the molecular level are exactly known, the system may nevertheless have so many degrees of freedom that it is even in principle impossible to obtain all the relevant information on the initial conditions that would be required in order to solve the dynamic equations for the entire system. Thus we believe that we know and understand the forces acting between atoms and

[1] The literature on music, in particular, contains plenty of jokes on this observation. As a chairman once said at a conference on signal analysis, ‘one man’s signal is another man’s noise.’

molecules; within classical mechanics it should be possible to describe the dynamics of, say, a cubic centimeter of gas to any degree of accuracy. However, this task would require that the initial positions of some 10^{17} atoms or molecules be known. We can safely assume that this is impossible, necessitating a statistical, or probabilistic, approach to the problem.

Even for systems with relatively *few* degrees of freedom, a somewhat similar situation may be encountered. In practice relevant initial conditions can be obtained only with a certain accuracy. The description of a wide class of physically interesting problems turns out to be extremely sensitive to the initial conditions, and the predicted temporal evolution is dramatically modified by even minute changes in these conditions. The resulting dynamic evolution can be described in statistical terms.

These few examples hint that an interpretation of statistical probabilities is that they represent systems regarding which we have insufficient *a priori* knowledge. For instance it can be argued that statistical mechanics assigns equal probabilities to all states with the same energy irrespective of the mechanical microstate simply because we do not have, and in practice *can* not have, information on these microstates. This principle of 'insufficient knowledge' can serve as a working hypothesis. It was apparently first formulated by Thomas Bayes (1763) stating that the absence of *a priori* knowledge can be expressed as *a priori* equal probabilities of events (Lee, 1989). This approach fails, however, in many respects, regarding quantum statistics in particular (Tolman, 1938, van Kampen, 1981). Mathematics can only derive the probabilities of outcomes of experiments or trials from *a priori given* probability densities or distributions. The only restrictions imposed on these probabilities are that they have to be positive definite and normalizable (and even this condition can formally be relaxed somewhat). There is no mathematical requirement that averages should exist, although it is sometimes hard to imagine describing a physical process without them.

- **Example:** The probability density

$$p_n(x) = n\frac{e^{-x}}{x} I_n(x)\,,$$

I_n being the modified Bessel function, is normalized for $0 < x < \infty$, i.e., $\int_0^\infty p_n(x)\,dx = 1$, for all $n = 1, 2, 3, \ldots$, but it has no average, i.e. $\int_0^\infty x p_n(x)\,dx$ diverges for any value of n.

In some cases symmetry arguments can help to determine the actual probability distributions, i.e. for a die we usually assume that the probability of each face coming up is $\frac{1}{6}$. In reality even this simple case relies on an idealization in terms of an absolutely exact cube with rounded corners. Even an honest die can strictly speaking never live up to this expectation, and the assumed probability is only an approximation. For slightly more complicated situations even our intuitive understanding of symmetry arguments is not quite straightforward, as the often-quoted example of Bertrand (1889) so elegantly demonstrates. He considered the problem of a straight line drawn at random to intersect a circle with unit radius, see Fig. 1.1. The question is then this: What is the probability, P, that the chord has a length longer than $\sqrt{3}$? The answer can be argued in three different ways, unfortunately giving three different results!

(1) Take a fixed point on the circle and consider all lines through this point assuming that the angle, θ, with the tangent is uniformly distributed in the interval $\{0; \pi\}$. All

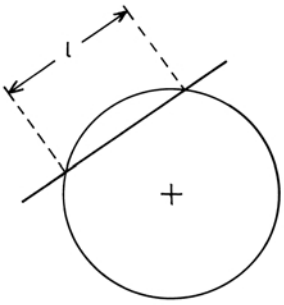

Figure 1.1 An illustration of Bertrand's problem with a circle with unit radius, where a chord is drawn at random. The chord length l is defined as the distance between the two points where the line intersects the circle.

lines cross the circle at two points except for the tangent itself, i.e. $\theta = 0$ and $\theta = \pi$, which on the other hand has a relative measure zero. In order for the chord to be longer than $\sqrt{3}$ it is required that $\pi/3 < \theta < 2\pi/3$. The interval is $\pi/3$, covering $\frac{1}{3}$ of the available interval of π, hence the answer is $P = \frac{1}{3}$.

(2) Consider all lines perpendicular to a fixed diameter of the circle. The chord is longer than $\sqrt{3}$ when the point of intersection lies on the middle half of the diameter. Consequently one finds $P = \frac{1}{2}$ by assuming the points to be uniformly distributed.

(3) For the chord to be longer than $\sqrt{3}$ its center must lie at a distance less than $\frac{1}{2}$ from the center. The area of a circle with radius $\frac{1}{2}$ is $\frac{1}{4}$ of the original circle. Assuming the chord centers to be uniformly distributed over the circle, the result $P = \frac{1}{4}$ is obtained.

The three examples are all based on the *a priori* assumption of uniform distributions, but of different quantities. Bertrand's question *has* an analytic answer, but only when it has been unambiguously formulated.

There is no obvious rescue for this and similar practical problems. There is no general method for obtaining unambiguous probability distributions for physical systems from first principles. In practice one can construct a hypothesis on the lowest possible level of description, follow its consequences for the statistical properties of the system, and eventually test the validity of the hypothesis against measurable quantities. A number of examples will illustrate this in the following chapters.

- **Exercise:** The following problem has little relevance for the present treatise, apart from giving an exercise in statistical reasoning. It is hoped that it can nevertheless give the reader some amusement before entering the more serious stuff!

The problem originates, at least in its present form, from the journal of the Danish Engineers Society, *Ingeniøren*. It goes as follows. Four gamblers, Colt, Browning, Smith, and Wesson reach a disagreement concerning a game of poker. The problem becomes so serious that it has to be settled by a gunfight. The four gentlemen meet at sunrise and agree that they should take turns to fire one shot at a time, to be continued until there is only one survivor. There seems after all to be some sense of fairness among them, so they decide that Mr Colt who is the best shot should come last as number four, since he hits every time he shoots. Mr Browning has a hit rate of 75%, so he comes as number three, Mr Smith has a record of hitting with 50% of his shots, so he is number two, while poor Mr Wesson, with a record of only 25%, shoots first. What is Mr Wesson's optimum strategy, and what is the probability of survival for each of the four gentlemen, assuming that each follows his optimum strategy?

Understandably, the winner decides to celebrate by throwing a big party. He has now many friends, as winners have, and invites 111 guests. To have a party in style, all the seats are labeled with the guests' names. Unfortunately, the first one to arrive does not notice, and takes a seat at random (i.e. he or she *might* take the correct seat). All the others take their own seats when they arrive, provided that they are available; otherwise they take seats at random. What is the probability that the last guest to arrive ends up in the seat with the correct label? Does the answer depend in any significant way on the number of seats being even or odd?

2 Elements of statistical analysis

Assume that we have a stochastic variable defined by a range of values, each associated with a certain probability. The variable is denoted X and the values it can assume x. The variable may assume discrete values only, e.g. it can be the number of particles in a volume element, each value, x_j, associated with a probability $P(x_j)$. Heuristically, we might interpret the probability of an event as the number of 'desired' events divided by the total number of relevant events. For simple cases like honest dies and card games, this interpretation is quite adequate, but it is insufficient as a general definition.

The sample space for relevant events may be discrete and finite as in Appendix A, or discrete and *in*finite as in Appendix B. The variable can be continuous such as, e.g., in Appendix C, for instance the voltage output of a noisy amplifier, with the probability of X having a value in a narrow range x, $x + dx$ being $P(x)\,dx$. Finally a combination of the two can occur, i.e. a mixture of discrete and continuous states as is encountered in atomic physics. The set of values, or *states*, that X can assume may be multidimensional, in which case it can conveniently be written as a vector $\boldsymbol{X}$. An example is the velocity components of a randomly moving particle.

By statistical averaging we understand the process based on the assumption that, ideally, infinitely many realizations are available. The average value of, say, X is then obtained by *ensemble averaging*:

$$\langle X \rangle = \lim_{N\to\infty} \frac{{}^1x + {}^2x + {}^3x + \cdots + {}^Nx}{N}, \tag{2.1}$$

where the indexes refer to the labels of individual realizations in the ensemble. If this is done in practice, only a finite number, N, of realizations is available, and only an *estimate*, i.e. an *approximation* to the actual average, can be achieved in this way. In terms of the normalized probability density $P(x)$, the mathematical process of averaging is expressed as

$$\langle X \rangle = \int_{-\infty}^{\infty} xP(x)\,dx, \tag{2.2}$$

for a continuous variable or, alternatively, in terms of a sum for discrete variables. It is evidently assumed that the probability density, $P(x)$, is known *a priori*. More generally, the average, or mean, of any quantity $f(X)$ is defined as $\langle f(X)\rangle = \int_{-\infty}^{\infty} f(x)P(x)\,dx$. Often we find the term *expectation value* $E\{f(X)\}$ for $\langle f(X)\rangle$. The term 'expectation value' might be somewhat misleading; assume that we have an honest die and calculate the average or expectation of the number $\mathcal{N}$ of dots. Since the numbers $\mathcal{N} = 1, 2, 3, 4, 5,$ and 6 are equally probable, we easily obtain $\langle \mathcal{N}\rangle = 3.5$. It is, however, unwise ever to actually *expect* the number 3.5 to come up, for only integer numbers can occur.

It may be important to note that, even though an event has zero probability, this does not necessarily mean that the event is physically impossible. The probability of an honest die showing 6 in *every* trial is zero, but there are no physical laws prohibiting this from actually happening.[1]

Probabilities can be expressed in terms of a *distribution function* or cumulative distribution function (Hogg and Craig, 1970), $\Pr(x)$, with the relation

$$\Pr(x) = \int_{-\infty}^{x} P(x')\,dx'$$

for the one-variable case. While $P(x)\,dx$ denotes the probability of finding x in the small interval $\{x, x+dx\}$, the distribution function $\Pr(x)$ gives the probability of finding x in the interval $\{-\infty, x\}$. For discrete variables the probability distribution is sometimes preferred, in order to avoid using the δ-functions (see Appendix D) which have to be introduced in the probability density for this case. The formulation in the following will be using only probability densities.

2.1 One-variable probabilities

First we discuss the case in which we are dealing with probabilities of one variable, say $P(x)$. Often we are not really interested in all the information contained in $P(x)$, but are content with averages such as $\langle X^n \rangle$, for some values of n. This information is more readily obtained, for instance, from moment-generating functions.

2.1.1 Generating functions

The average of $\exp(\alpha x)$, called the *moment-generating function*, is defined as (Bendat, 1958, Hogg and Craig, 1970)

$$M(\alpha) \equiv \langle \exp(\alpha X) \rangle = \int_{-\infty}^{\infty} e^{\alpha x} P(x)\,dx. \tag{2.3}$$

In particular we find by using (2.3) the derivatives

$$\frac{d^n M(\alpha)}{d\alpha^n} = \int_{-\infty}^{\infty} x^n e^{\alpha x} P(x)\,dx.$$

[1] It is often said that it does not make sense to discuss the probability of events that *have* already occurred, and of course much can be said in justification of this statement. However, even proponents of this point of view will probably (like the author and most readers) start wondering if the opponent in a die game continuously gets 6 or whatever number is needed to win. Faced with an event that has already happened, it is justifiable to seek its explanation by considering the probabilities of various causes for its occurrence. When different explanations for an event are possible, giving preference to the cause with the highest probability is a fully acceptable procedure. Maximum-likelihood methods (Hogg and Craig, 1970) are based on this basic hypothesis, by virtue of the assumption that it is the most probable event which is actually being observed.

The moment-generating function is evidently a function of the variable α and serves to generate the averages, or *moments*, of X, i.e. $\langle X \rangle$, $\langle X^2 \rangle$, etc. which are obtained by evaluating $dM(\alpha)/d\alpha$, $d^2M(\alpha)/d\alpha^2$, etc., at $\alpha = 0$. The individual terms in a Taylor expansion of $M(\alpha)$ will consequently contain the averages $\langle X^n \rangle$ in increasing order. However, not every probability density has an associated moment-generating function, and it is often advantageous to consider the *characteristic function* defined as the Fourier transform of the probability density

$$CH(\alpha) \equiv \langle \exp(i\alpha X) \rangle = \int_{-\infty}^{\infty} e^{i\alpha x} P(x)\, dx. \tag{2.4}$$

With rather mild restrictions on the variation of $P(x)$ at large $|x|$, the moments of X are then given by

$$\langle X^n \rangle = (-i)^n \frac{d^n}{d\alpha^n} CH(\alpha) \bigg|_{\alpha=0}, \tag{2.5}$$

the subscript indicating that the derivative is to be taken at $\alpha = 0$. For a discrete variable, the characteristic functions

$$CH(\alpha) \equiv \sum_{j}^{\infty} P(x_j) e^{i\alpha x_j} \tag{2.6}$$

can be introduced, where $P(x_j)$ is the probability of the jth event. Here, the characteristic function is a sum of exponentials.

For the case in which a variable n takes on only integer values, a definition in terms of the z-transform (Oppenheim and Schafer, 1975) gives the generating function

$$\Gamma(z) \equiv \langle Z^n \rangle = \sum_{n=-\infty}^{\infty} z^n P_n. \tag{2.7}$$

Here n is a *lattice-type* random variable. We recognize $\Gamma(1/z)$ as the z-transform of P_n with n taking on only integer values (Papoulis, 1991). (The Γ-function introduced here should not be confused with the gamma-function which interpolates $n!$) On differentiating (2.7) k times we obtain

$$d^k\Gamma(z)/dz^k = \langle n(n-1)\cdots(n-k+1)z^{n-k} \rangle,$$

which for $z = 1$ becomes $\langle n(n-1)\cdots(n-k+1) \rangle$.

- **Exercise:** Derive the moment-generating function for a Poisson distribution.
- **Exercise:** Obtain the moment-generating function and the characteristic function for the binomial distribution.
- **Exercise:** Derive the moment-generating function for a Gaussian distribution $P(x) = (2\pi A)^{-1/2} \exp(-\frac{1}{2}x^2/A)$ and demonstrate that $\langle X^4 \rangle = 3\langle X^2 \rangle^2$.
- **Exercise:** Demonstrate that the characteristic function for a variable Z that is the sum of two independent random variables, $Z = X + Y$, is the product of the individual characteristic functions. Generalize this result to a sum of arbitrarily many independent variables.
- **Example:** By a random vector $\boldsymbol{L}$, we understand a vector drawn in a random direction and possibly with length, L, chosen at random also (Feller, 1971). For

a vector of *unit* length in three dimensions, $\mathcal{R}^3$, the distribution of the projection, L_x, on an axis is uniform over $\{0, 1\}$. The projection of the same vector on a *plane* has a probability density $\ell/\sqrt{1-\ell^2}$ for $0 < \ell < 1$.

It is important to note that the result depends on the dimensions of the problem. For a vector of unit length in two dimensions, $\mathcal{R}^2$, the distribution of the projection, L_x, on an axis is $2/(\pi\sqrt{1-\ell^2})$ for $0 < \ell < 1$.

Consider now the sum of two independent random unit vectors in $\mathcal{R}^2$. The resultant of these vectors has a length L, with a probability density $2/(\pi\sqrt{4-\ell^2})$ for $0 < \ell < 2$. Actually, by the law of cosines, $L = \sqrt{2 - 2\cos\gamma} = |2\sin\left(\frac{1}{2}\gamma\right)|$, where γ is the angle between the two vectors, and $\frac{1}{2}\gamma$ is by assumption uniformly distributed in $\{0, \pi\}$.

2.1.2 Changes of variables

An important question concerns the change in variables; often we know *a priori* the probability density for a certain event, and want to determine the probability density for something else that depends in a deterministic way on the outcome of this event. As a simple example, we may consider X to be a temperature and Y to be the length of a metal rod, which is varying due to thermal expansion in some, possibly nonlinear, way. Assume that the probability density, $P_X(x)$, of the event X is known, and that another event Y is a deterministic consequence of X, i.e. $Y = f(X)$. It is then, at least in principle, straightforward to determine the probability density $P_Y(y)$. The probability that Y has a value in the range $\{y; y + \Delta y\}$ is

$$P_Y(y) = \int \delta[f(x) - y] P_X(x)\, dx\,. \tag{2.8}$$

In the case in which there is a one-to-one correspondence between X and Y, the transformation is readily expressed as

$$P_Y(y)\, dy = P_X(x)\, dx. \tag{2.9}$$

In case there are problems with the sign (probabilities had better be positive numbers!) this expression is to be interpreted as $P_Y(y) = P_X(x)|J|$, J being the Jacobian determinant for the general multivariable case. The expression (2.9) can be most useful.

- **Exercise:** Assume that a circle with radius R is placed randomly with its center on the x-axis, with positions uniformly distributed in the interval $\{0; \mathcal{L}\}$. Determine the probability density of chord lengths along the y-axis, i.e. the distribution of the lengths of segments of the y-axis inside the circle. Let the ratio $R/\mathcal{L}$ be arbitrary. Repeat the problem, now with a *sphere* with radius R placed randomly in the x–z plane, with its center uniformly distributed in a square $\{0; \mathcal{L}\}, \{0; \mathcal{L}\}$.
- **Exercise:** Assume that the variable X has a Gaussian distribution, $P(x) = (2\pi\sigma^2)^{-1/2}\exp(-\frac{1}{2}x^2/\sigma^2)$. Consider the variable $Y = X^2$ and demonstrate that its probability density is given by the *chi-square* probability density

$$P(y) = \frac{1}{\sigma\sqrt{2\pi y}}\exp\left(-\frac{y}{2\sigma^2}\right), \tag{2.10}$$

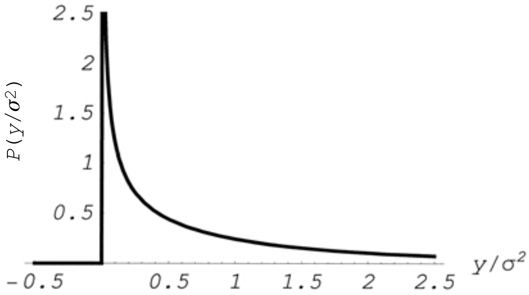

Figure 2.1 The normalized chi-square distribution, $P(y/\sigma^2)$, as given by (2.10).

for $y \geq 0$ and $P(y) = 0$ otherwise, see Fig. 2.1. Demonstrate that $\langle Y \rangle = \sigma^2$ and $\langle Y^2 \rangle = 3\sigma^4$. Prove explicitly that $P(y)$ is indeed normalized, $\int_0^\infty P(y)\, dy = 1$.

- **Exercise:** Assume that a variable X has a Gaussian distribution,

  $P(x) = (2\pi\sigma^2)^{-1/2} \exp(-\frac{1}{2}x^2/\sigma^2)$.

  Consider another variable

  $Y = (2/\pi)^{-1/2} \int_0^X \exp(-\frac{1}{2}\gamma^2)\, d\gamma$ for $X \geq 0$,

  while $Y = 0$ for $X < 0$. Derive the probability density for Y.

2.2 Multivariable probabilities

The random variable can be multidimensional, as has already been mentioned, i.e. the corresponding probabilities can depend on many different variables. Averages are defined by a simple generalization of (2.2) as

$$\langle f(X_1, X_2, \ldots, X_N)\rangle = \int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_N) P(x_1, x_2, \ldots, x_N)\, dx_1\, dx_2 \ldots dx_N, \quad (2.11)$$

where $P(x_1, x_2, \ldots, x_N)$ is the joint probability density for the variables $x_1, x_2, \ldots, x_N$.

For statistically independent variables, their joint probability density is the product of their individual probability densities, e.g.

$$P(x_1,\ x_2, \ldots, x_n) = P(x_1)P(x_2) \ldots P(x_N).$$

Pairwise independence does not ensure absulute independence; assume for instance $P(x_1, x_2) = P(x_1)P(x_2)$ and $P(x_2,\ x_3) = P(x_2)P(x_3)$, but this does not mean that x_1 and x_3 are independent, i.e. $P(x_1,\ x_3) \neq P(x_1)P(x_3)$ in general. The proof is trivial.

Moment-generating functions and characteristic functions can be defined for multivariate distributions as well. For instance, for a bivariate probability density, $P(x_1, x_2)$, we have

$$CH(\alpha, \xi) \equiv \langle \exp(i\alpha X_1 + i\xi X_2)\rangle$$
$$= \int_{-\infty}^{\infty} e^{i\alpha x_1 + i\xi x_2} P(x_1, x_2)\, dx_1\, dx_2. \quad (2.12)$$

Averages or moments are then given by

$$\langle X_1^n X_2^m \rangle = (-i)^{n+m} \frac{d^{n+m}}{d\alpha^n d\xi^m} CH(\alpha, \xi)\bigg|_{\alpha=\xi=0} . \tag{2.13}$$

A probability can be *conditional*, with $P(x\,|\,y)$ giving the probability for the event x, given with certainty the event y. Then, $P(x) = \int_{-\infty}^{\infty} P(x\,|\,y)\,dy$ for a continuous variable, or a corresponding sum for a discrete variable. By Bayes' rule we have

$$P(x\,|\,y) = \frac{P(x, y)}{P(y)}, \tag{2.14}$$

where $P(x, y)$ is the joint probability for x and y. Bayes' rule has self-evident generalizations to multivariable probabilities. It is often easier to predict or argue expressions for *conditional* probabilities than it is for their *un*conditional counterparts. The inverted version of (2.14) can then sometimes be used to obtain the full joint probability.

- **Examples:** A harmonic oscillation $\tau_1 \cos(t + \tau_2)$ with a random phase τ_2 and a random amplitude τ_1 represents a sample function with two parameters. A function that alternates between $+1$ and -1 at N random times $t = \tau_n$ with $n = 1, 2, \ldots, N$ represents a sample function with N parameters.

2.2.1 Correlation

The *covariance* of two random variables X and Y can be defined as

$$\begin{aligned} Cov &\equiv \langle (X - \langle X \rangle)(Y - \langle Y \rangle) \rangle \\ &\equiv \langle XY \rangle - \langle X \rangle \langle Y \rangle . \end{aligned} \tag{2.15}$$

The *correlation coefficient* for the relationship between two random variables X and Y can be defined as

$$C \equiv \frac{\langle (X - \langle X \rangle)(Y - \langle Y \rangle) \rangle}{\sqrt{\langle (X - \langle X \rangle)^2 \rangle \langle (Y - \langle Y \rangle)^2 \rangle}} . \tag{2.16}$$

For the case $X = Y$ we have $C = 1$, identically. If, on the other hand, X and Y are independent variables, we readily find that $C = 0$. For complex variables, we define $Cov \equiv \langle XY^* \rangle - \langle X \rangle \langle Y^* \rangle$, where the asterisk denotes the complex conjugate.

2.3 Stochastic processes

Consider a measurable quantity $Y(t)$ that is a function of some variable t. Let $Y_x(t)$ denote an ensemble of such functions of t, labeled by a random parameter x with a given distribution, $P(x)$. When t stands for time it is customary to call $Y(t)$ a random or stochastic process. A *sample function* or *a realization* of the process is obtained for one particular value of x. An ensemble of such sample functions thus constitutes a stochastic process. The averaging is understood as an ensemble averaging in the sense indicated in (2.1).

- **Examples:** A particularly simple example of a stochastic process is one in which each sample function is for all times a constant x that varies over the ensemble. A