6

Distributions

6.1 Introductory Remarks

6.1.1. Thus far, we have been occupied with the conceptual aspects of the formulation, and the thoroughness of the treatment reflects what the material seemed to us to require. Likewise, we have chosen to deal with the simplest topics and problems, whose meaning was not obscured by the need to involve complicated mathematics (but in fact contributed by appearing in a clear and simple light).

The time has now come, however, to abandon these self-imposed limitations. We must examine whether, and to what extent, we can implement, in any domain whatsoever, the study of probability in terms of the image most often thought of (in an informal manner); that of the 'distribution of mass'. In actual fact, of course, it is well known that the notion of a probability distribution (the precise mathematical translation of this image) is taken directly as the starting point in many approaches, particularly modern ones. The aim of the present chapter is to introduce this notion and the requisite mathematical tools, tying them in rigorously with our previous formulation and making any necessary modifications or limitations.

There are, therefore, two different aims to bear in mind in what follows: on the one hand, to provide a knowledge of the mathematical tools required in further study of the calculus of probability; on the other hand, to give the mathematical and conceptual details which derive from our previously established formulation and point of view.

6.1.2. We shall try to satisfy the first aim as concisely as possible, quoting, with a minimum of explanation, and without proof, those things that can be found in any book on probability, or whose proof can be obtained either with a standard knowledge of analysis or on an intuitive basis. Alternatively, if the reader wishes, the proofs can be taken for granted and this will not affect applications or further reading.

6.1.3. Our second aim, one of a critical nature, will need a more careful treatment, at greater length. Although we do not wish to dwell upon it more than we have to, any omission or incompleteness in what is necessary would certainly cause misunderstanding and incomprehension (especially among those readers who, by interpreting certain sentences in the standard way, would find them, and quite rightly so, either incomprehensible, or, misunderstanding them, wrong). For this reason, we strongly recommend the reader, and especially those who think that they already know enough

Theory of Probability: A Critical Introductory Treatment, First Edition. Bruno de Finetti.

about the topic of this chapter, not to skip it, and to dwell, in particular, upon the details relating to the differences, slight but important, between this and the standard interpretations.[1]

6.2 What we Mean by a 'Distribution'

6.2.1. An abstract and general explanation would, at this stage, appear rather vague and colourless. It is more appropriate to consider here the simplest and most important special case, that of distributions on the real line, together with their various interpretations. These interpretations should all be kept in mind, in order that the most convenient one can be called upon in any particular instance. This special case will eventually be revealed to have a relationship with that of random quantities in general.

Proceeding in the usual way, we introduce immediately, as a starting point, and as the main mathematical tool for the definition of a distribution, a function $F(x)$, increasing[2] from 0 (as $x \to -\infty$) to 1 (as $x \to +\infty$), and called a *distribution function*.

6.2.2. As a *first interpretation*, the most intuitive one, we have that of a *distribution of mass* on the real line (with the assumption that 'total mass' = 1). $F(x)$ is the mass to the left of a point x, $1 - F(x)$ the mass to the right; the increment $F(x'') - F(x')$ is the mass in the interval $x' \leqslant x \leqslant x''$. If there is a mass, p_h, *concentrated* at the point x_h, F is discontinuous at x_h and p_h is its 'jump', $F(x_h + 0) - F(x_h - 0)$.[3] There is at most a finite or countable number of such jumps, and F is continuous elsewhere.

A distribution that only has concentrated masses ($\Sigma_h p_h = 1$) is called *discrete*; one without concentrated masses is called *continuous*. The most familiar case of the latter is that of *absolutely continuous* distributions; those admitting a *density* function, $f(x) = F'(x)$, such that

$$F(x) = \int_{-\infty}^{x} f(x)\ \mathrm{d}x.$$

In actual fact, when the term 'continuous' is used, it is this special case which is often understood. There is, however, an intermediate case between the *discrete* and *absolutely continuous;* that of *continuous but not absolutely continuous*. In 6.2.3 we shall make this idea concrete by means of an example (and this example will also have an interesting interpretation in a problem in probability). For the time being, we shall limit ourselves to the definition and the basic properties.

6.2.3. To say that $F(x)$ is continuous means, as everyone knows, that for each ε, however small, every interval whose length is less than some suitable δ contains a

---

1 Recall the warnings given already in Chapter 1 (1.2.1).

2 We use 'increasing' to mean 'nondecreasing'; we shall use 'strictly increasing' if the function is not constant in any interval.

3 These two values must be distinguished when considering $F(x)$ if there is a jump at the point x (and we have a choice according on whether the mass at x is to be considered together with those on the left or those on the right). For various reasons (see 6.5.1), we prefer to avoid those conventions which make $F(x)$ one-to-one at the discontinuity points (by saying that it assumes *all* the values y, $F(x - 0) \leqslant y \leqslant F(x + 0)$. However, when dealing with statistical distributions, where some convention is necessary, we shall take $F(x) = F(x + 0)$ (as is necessary if 'individuals with h children' is to mean 'including those with exactly h children').

We apologize for the awful notation $F(x + 0)$; it is, however, concise and unambiguous.

mass $<\varepsilon$. To say that it is *absolutely continuous* (Vitali) means something more: that the same is true of the mass contained in any arbitrary number of intervals of total length less than δ.[4]

Every distribution $F(x)$ can be decomposed into partial distributions of masses of the three types. We first of all set

$$F(x) = a_C F_C(x) + a_B F_B(x) + a_A F_A(x) \quad (a_C + a_B + a_A = 1)^5 \tag{6.1}$$

where:

$a_C = \sum_h p_h$ is the sum of the concentrated masses (masses of type C),
$a_C F_C(x) = \sum_h p_h (x_h \leqslant x)$ is the sum of these masses in $[-\infty, x]$.

We now consider the residual partial distribution,

$$F_{AB}(x) = F(x) - a_C F_C(x),$$

that is, $F(x)$ without the concentrated masses; it follows that:

a_B = 'total mass of type B' = upper limit of the mass of $F_{AB}(x)$ which can be enclosed within intervals of arbitrarily small total length,
$a_B F_B(x)$ = total mass of type B in $[-\infty, x]$ (detailed definition as above).

We are left with $a_A F_A(x) = F(x) - a_C F_C(x) - a_B F_B(x)$, and this is the absolutely continuous part of the distribution (the masses of the first two types, which do not fulfill the condition of absolute continuity, having been removed).

It is easy to see that, in a linear combination of distributions,

$$F(x) = c_1 F_1(x) + c_2 F_2(x) \quad (c_1 + c_2 = 1),$$

the various types are preserved. It follows, therefore, that the F_C, F_B, F_A of an arbitrary linear combination are the linear combinations of the corresponding parts of the summands (in particular, a particular type of mass exists in the linear combination if and only if it exists in at least one of the summands). If we say that a distribution is of type A, B, C, AB, AC, BC, ABC, to indicate the pure types involved in it, we can express our conclusion by saying that in a linear combination the letters of the types combine (e.g. from AC and BC we get ABC).

An example of a type B distribution. The following procedure can be used to construct the well-known Cantor set (of measure zero, even in the Jordan–Peano sense) and a distribution on it (which is therefore of type B).

Let us divide the interval [0, 1] into three equal parts. In the middle interval, $[\frac{1}{3}, \frac{2}{3}]$, we set $F(x) = \frac{2}{3}$, so that no mass is placed there, and half the mass is placed in each of the first and third intervals. This operation is then repeated in these latter two intervals. In

4 It makes no difference whether we consider the number of intervals as *finite* or infinite (countable: it cannot be uncountable). It is understood that $\varepsilon > 0$ and $\delta > 0$.

5 Obviously, if $a_i = 0$ (i.e. one of the components is missing) the corresponding F_i is missing. The meanings of the letters are: C = concentrated; A = absolutely continuous; B = intermediate case between C and A.

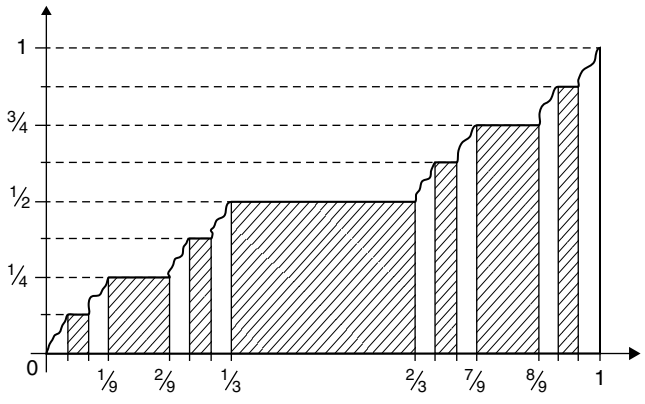

Figure 6.1 The Cantor distribution.

each of them we will have three subintervals (of length $\left(\frac{1}{3}\right)^2 = \frac{1}{9}$), and we set $F(x) = \frac{1}{4}$ (respectively, $= \frac{3}{4}$) on the central intervals, thus excluding masses there. The mass is then placed on the four residual intervals, $\frac{1}{4}$ on each.

Proceeding in this manner (Figure 6.1), after n steps $F(x)$ is defined (with values which are multiples of $\left(\frac{1}{2}\right)^n$) on the whole interval [0, 1], except for the 2^n residual parts, each of length $\left(\frac{1}{3}\right)^n$, where all the mass resides ($\left(\frac{1}{2}\right)^n$ on each residual interval). In the limit, $F(x)$ is defined everywhere and is continuous. It is not, however, absolutely continuous: after n steps the mass is contained within the 2^n intervals each of length $\left(\frac{1}{3}\right)^n$, and $\left(\frac{2}{3}\right)^n$ in total. It can, therefore, be contained within a finite number of intervals of total length less than any given $\varepsilon > 0$.

A probabilistic interpretation. It might be thought that the above construction merely serves to provide a critical comment; giving a pathological example with no practical meaning. On the contrary, we can give a simple practical example of a problem in probability where such a distribution arises.

Suppose we wish to pick a real number in [0, 1] by successively drawing from an urn the digits of its decimal representation:

$$X = 0 \bullet X_1 X_2 X_3 \ldots X_n \ldots, \quad \text{i.e.} \quad X = \sum X_n / B^n \left(B = \text{base; e.g. } 10 \right).$$

If a ball representing a figure is missing, all the numbers containing it become impossible (i.e. some intervals are excluded, as in the example given). The above example corresponds to the assumption that $B = 3$, with the figure 1 missing (only the numbers with 0 and 2 are possible, like 0·22020002020022202….).

It is rather surprising to note that this happens even if the balls are all present (unless all of them have the same probability $1/B$).[6] If one of the figures has probability $p < 1/B$, and we take c between p and $1/B$, and N sufficiently large, the set of numbers X in which

6 This observation is too obvious to be novel; however, I do not remember having seen it before, and I had not thought of it prior to adding it here to the usual example.

that figure appears in the first N places with frequency $\geqslant$c has measure arbitrarily close to 1 and mass arbitrarily close to 0.[7]

6.2.4. Let us observe now how a different interpretation of F permits us to extend considerably its applicability and effectiveness. Given any interval I (with extreme points x' and x''), it suffices to set $F(I) = F(x'') - F(x')$ to obtain F as an additive function for the intervals. If we identify the intervals with their indicator functions ($I(x) = (x' \leqslant x \leqslant x'') = 1$ or 0 depending on whether x belongs to I or not), we obtain F as a linear functional, defined for every $\gamma(x) = \Sigma_h y_h I_h$ (step functions with values y_h on the disjoint intervals I_h) by $F(\gamma) = \Sigma_h y_h F(I_h)$. This can be extended to all functions $\gamma(x)$ which can be approximated, in an appropriate way, from above or below, by means of step functions. More precisely, $F(\gamma)$ is determined if, thinking of γ' and γ'' as generic step functions such that

$$\gamma'(x) \leqslant \gamma(x) \leqslant \gamma''(x)$$

everywhere, we have $\sup F(\gamma') = \inf F(\gamma'')$, hence $F(\gamma)$ necessarily has that same value since $\sup F(\gamma') \leqslant F(\gamma) \leqslant \inf F(\gamma'')$.

In actual fact, what we have defined, in a direct and somewhat abstract way, is nothing other than the integral

$$\phi(\gamma) = \int \gamma(x)\, \mathrm{d}F(x) = \int \gamma(x) f(x) \mathrm{d}x \quad \left(\int \text{represents} \int_{-\infty}^{+\infty} \right), \tag{6.2}$$

where the first expression (one which always holds) is the Riemann–Stieltjes integral, and the second (which only holds for absolutely continuous distributions) is the Riemann integral.

As an example, suppose we consider the two functions

$$\gamma(x) = x = \square(x) \quad \text{and} \quad \gamma(x) = x^2 = \square^2(x).$$

In this case, $F(\square)$ = the abscissa of the barycentre, and $F(\S^2)$ the moment of inertia (about the origin) of the mass distribution. In integral form,

$$F(\square) = \int x\, \mathrm{d}F(x) = \int x f(x) \mathrm{d}x, \qquad F(\square^2) = \int x^2\, \mathrm{d}F(x) = \int x^2 f(x)\, \mathrm{d}x.$$

As possible interpretations of the function, $\gamma(x)$, one might, for instance, think of it as representing (for the mass at x) the reciprocal of the density, or the percentage by weight of a given component (e.g. of a given metal if we are dealing with an alloy whose composition varies with x), or the (absolute) temperature. In these three cases, the integral, apart from constant terms, will yield the total volume, the weight of the given component, and the quantity of heat, respectively.

6.2.5. A *second interpretation* is the statistical one. It is convenient to mention it here in order to draw attention to the practical importance of the notion of distribution in the field of statistics. This is not only closely connected and related to the probabilistic notion but also provides it with problems and applications. However, we shall reserve discussion of this until later.

---

7 This assertion will be seen as obvious as soon as we encounter the basic ideas of 'laws of large numbers' (Chapter 7, Section 5).

In the final analysis, the image is the same as before: that of a mass distribution. In fact, the distribution of a population of n individuals, on the basis of any quantitative characteristic whatsoever, can be thought of as obtained, in the case of number of children, for example, by placing a mass $1/n$ at the point $x = h$ for each individual with h children (h = 0, 1, 2,…), or, in the case of height, at the points $x = x_i$ (distinct if the measurements are sufficiently precise), denoting by i = 1, 2, …, n the n individuals, and by x_i their heights.

In the first example, we have masses $p_h = n_h/n$ concentrated at the points $x = h$ (n_h denotes the number of individuals with h children) and therefore:

$$F(x) = \sum_h (n_h / n)\left(h \leqslant x\right)$$
$$= \text{the percentage of individuals with not more than } x \text{ children.}$$

In the second example (let us assume that the individuals have been indexed in order of increasing height), we have a jump of $1/n$ at each point x_i (and, if n were large, one could in practice consider the distribution to be continuous – if necessary by 'smoothing'), and the distribution function is given by

$$F(x) = (1/n)\ \max\ i\ \left(x_i \leqslant x\right) \quad \left(\text{that is: } F(x) = i/n,\ \text{for } x_i \leqslant x \leqslant x_{i+1}\right).$$

Alternatively, one might be interested in performing some kind of 'weighting' instead of simply 'counting' the individuals (for example: instead of $1/n$ throughout, a 'weight' might be chosen proportional to income, average number of bus journeys per day, cups of coffee consumed, etc., depending on what was of interest). The 'population' might consist of objects, or events, or anything but it is customary to retain the terms 'population' and 'individuals'. If a generic and neutral term is required, one can use 'statistical units'. In the general case, units may be counted straightforwardly, or with some appropriate 'weighting'.

This will suffice for the present. We merely recall (see Chapter 5, Sections 5.8–5.10) that a statistical distribution *is not* a probability distribution, although it can, in various ways, give rise to one.

6.2.6. In order to clarify, from a different angle, certain aspects of the above (and, more importantly, to mention some further extensions) it is useful at this point to introduce a *third interpretation*. An additive function (non-negative, and with its maximum = 1) is also called a *measure*; the change in nomenclature, from mass to measure, is of no importance, but the fact that we have at hand a natural way of looking at such a 'measure' – or, to be precise, the 'F-measure' – in terms of its own scale (of length) is important.

One works in terms of this scale by looking at $y = F(x)$ instead of at x (as was clear from the definition). We have only to observe that, by drawing the graph of the distribution function (Figure 6.2a), we establish an (ordered) correspondence between the points of the x-axis (all of it) and those of the interval [0, 1]. The mass of any arbitrary interval on the x-axis is then measured by the length of its image on the y-axis. Of course, the correspondence is not necessarily one to one (it will be so if $F(x)$ is strictly increasing from $-\infty$ to $+\infty$). To a point of discontinuity on the x-axis there corresponds, on the y-axis, an interval whose length equals the mass which is concentrated at that point; to any interval of the x-axis on which $F(x)$ is constant (no mass) there

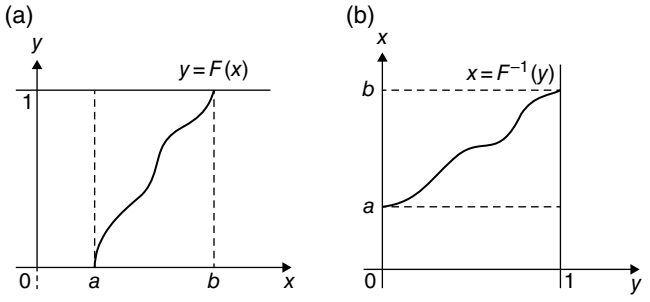

Figure 6.2 The graphs of (a) the distribution function and (b) its inverse: $y = F(x)$ and $x = F^{-1}(y)$.
In addition to the present (measure theoretic) interpretation, we have also seen that the statistical interpretation, as *graph of the distribution*, is of interest (and is the most useful from the point of view of applications). In Section 6.4 we shall further consider the probabilistic setting, in which the above admits the following interpretation: one can always construct a random quantity with a preassigned distribution F starting from a Y with a uniform distribution on [0, 1] (or, conversely, $Y = F(x)$ has a uniform distribution on [0, 1] if X has distribution F; some device is necessary in order to make it uniform at the jumps).

corresponds a single point of the y-axis. Apart from the interpretation in mechanical terms, this is also clear geometrically. Observe that in both cases the graph $y = F(x)$ (conveniently thought of as containing, for discontinuity points x, all the y's between $F(x \pm 0)$) contains, respectively, vertical or horizontal segments which project to a single point of the orthogonal axis.

In order to concentrate attention on the measure, and to make it easier to visualize developments based upon it, we find it convenient to reverse the rôles of the x- and y-axes, and to look instead at the graph of $x = F^{-1}(y)$ (Figure 6.2b).

Note that the change of variable from x to y transforms, for example, the Stieltjes integrals into ordinary integrals:

$$F(\gamma) = \int \gamma(x)\,\mathrm{d}F(x) = \int \gamma\left(F^{-1}(y)\right)\mathrm{d}y = \int \gamma(x)\,\mathrm{d}y.$$

6.2.7. We shall see later that this form of representation is also useful for visualizing many problems and situations in the theory or probability and statistics (see, for example, Section 6.6). What is of immediate interest, however, is to exploit *the fact that we have, on the y-axis, the F-measure 'on its natural scale'* in order to look, succinctly, and without formulae, at the question of possible further extensions.

In terms of y, $F(\gamma)$ corresponds to the ordinary (Riemann) integral, and therefore $F(I)$ (where I = set, thought of as identified with its indicator function, $I(x) = (x \in I)$) can be interpreted as the Jordan–Peano measure of the image set of I on the y-axis. The F-measure (apart from the given transformation) is the *J–P measure* (that is, Jordan–Peano), and the *F-measurable* sets are those whose image, on the y-axis, is a *J–P-measurable* set.

6.3 The Parting of the Ways

6.3.1. At this point we are faced with a choice.

It is well known that *there exists a unique extension of the J–P measure to a much larger class of sets*. The methods used are due to Borel and Lebesgue, and the basic idea

(put rather crudely) is to argue about a countably infinite collection of sets as if they were a finite collection; in particular, one invokes countable additivity (valid not only for the sum in the ordinary sense, but also for the sum of convergent series). Similar considerations apply to the extension of the notion of integral.

From the viewpoint of the pure mathematician – who is not concerned with the question of how a given definition relates to the exigencies of the application, or to anything outside the mathematics – the choice is merely one of mathematical convenience and elegance. Now there is no doubt at all that the availability of limiting operations under the minimum number of restrictions is the mathematician's ideal. Amongst their other exploits, the great mathematicians of the nineteenth century made wise use of such operations in finding exact results involving sums of divergent series: first-year students often inadvertently assume the legitimacy of such operations and fail the examination when they imitate these exploits. At the beginning of this century it was discovered that there was a large area in which the legitimacy of these limiting operations could be assumed without fear of contradictions, or of failing examinations: it is not surprising therefore that the tide of euphoria is now at its height. Two quotations, chosen at random, will suffice to illustrate this[8]: 'The definition is therefore *justified ultimately by the elegance and usefulness* of the theory which results from it'; 'Conditions about the continuity of (the integral) are really essential if the operation is to be a useful tool in analysis – *there would not be much of analysis left* if one could not carry out at least sequential limiting operations'.

6.3.2. Are there any reasons for objecting to this from a mathematical standpoint? Rather than 'objections', I think it would be more accurate to speak of 'reservations'; there are, I believe, two reasons for such reservations.

The first concerns what happens *outside* of that special field which results from the above approach. It has been proved (by Vitali, and afterwards, in more general contexts, by Banach, Kuratowski and Ulam) that if one is not content with finite additivity, but insists on countable additivity, then it is no longer possible to extend the 'measure' to *all the sets* (whereas there is nothing to prevent the extension to all sets of a finitely additive function which coincides – when they exist – with the J–P measure, or the L-measure).

Countable additivity cannot, therefore, be conceived of as a general principle which leads us safely around *within* the special field, and allows us to roam *outside*, albeit in an undirected manner, with an infinite number of choices. On the contrary, it is like a good-luck charm which works *inside* the field, but which, on stepping outside, becomes an evil geni, leading us into a labyrinth with no way out.[9]

---

8 From J.F.C. Kingman and S.J. Taylor, *Introduction to Measure and Probability*, Cambridge University Press (1966), pp. 75 and 101 (the italics are mine).

9 This image of a labyrinth with no way out is an exact description of the situation. In fact, if one wishes to extend the definition of L-measure to nonmeasureable sets, respecting countable additivity, this can always be done step by step (choosing, for a given set, a value at random between the two extremes of inner and outer measure as determined by the extension so far made). After an infinite number of steps, however, a contradiction can arise, and sooner or later (before exhausting *all* the sets) it certainly arises. (As an analogy; a convergent series remains such if we add 1 onto a finite number of terms, no matter how far we go … , but not if we add 1 onto all the terms!)

This observation renders even more artificial the distinction between those sets which are L-measurable and those which are not (none of them has any particular feature which makes it unsuitable).

Never mind, it might be argued: measurable sets will suffice. But from what point of view? Practically speaking, the intervals themselves were perhaps sufficient. From a theoretical standpoint, however, is there any justification for this discrimination between sets of different *status;* the orthodox which we are permitted to consider, and the heretical which must be avoided at all costs? Would it be too far-fetched to suggest an analogy with real numbers, some of which still bear the name irrational because their existence had so scandalized the Pythagoreans?

6.3.3. The second reason for the reservation spoken of earlier concerns what happens *inside* the special field. Here the rules are more restrictive and permit us only to follow a uniquely defined path – like a runway for an automatic landing. This may be a fine thing but must one be compelled to invoke this aid in all possible cases? Is it too absurd to believe that soap may sometimes have its uses despite the existence of an infinite number of detergents, each of which washes infinitely whiter than any of the others?

We can, happily, provide mathematical analogies in this case, and these will be more illuminating than the whimsical variety (although the latter may help in suggesting in advance the sense of the mathematics).

First a trivial example: if the value of a function is given at a finite number of points (or at a countably infinite number) I can complete it in an infinite number of ways – even under additional conditions (like continuity, etc.). Given n values, I know that the problem has one and only one solution if I add the condition that the function is a polynomial of degree $n - 1$ (if $n = \infty$, and I add the condition that the function be analytic, there is either one solution or no solution): is this a good reason for limiting oneself to this particular solution; or for considering it as 'special'?

A further example seems to me rather relevant. There exist methods for summing series – for example, that of Cesàho – which often give a uniquely determined answer in cases where the usual method of summation leads only to (different) upper and lower limits. Is it right that as a result we should always interpret 'sum of a series' as meaning Cesàro sum, and to banish as 'outmoded' the usual notion of convergence? Of course, the compass of the Cesàro procedure (even iterated) is not comparable to that of other innovations, like that of Lebesgue, but, even assuming it to be such, would it then be justified? And would it not be possible that in certain cases there would be interest in ascertaining whether, in fact, the series were convergent according to the old definition (which, although out of date, has not become meaningless)? What if we wanted to know, in that sense, the upper and lower limits? In my opinion, this example is apposite in every respect. In the case of Lebesgue measure, as for Cesàro summability, there is a procedure that (because of additional conditions) often yields a unique answer instead of bounding it inside an interval within which it is not determined. Whether one solution is more useful than the other depends on further analysis, which should be done case by case, motivated by issues of substance, and not – as I confess to having the impression – by a preconceived preference for that which yields a unique and elegant answer *even when the exact answer should instead be 'any value lying between these limits'.*

6.3.4. The above remarks, made from a purely mathematical standpoint, are not designed to prove anything other than that the case for consigning the Riemann integral to the attic now that the Lebesgue integral is available has not itself been proved. The Riemann integral can still be a necessary tool; *not in spite of* its indeterminacy, but *precisely because of* it: this indeterminacy may very well have an essential meaning.

On the other hand, from a mathematical point of view I would by no means presume to discuss topics in analysis which I only know in the context of what I need. In the case of the calculus of probability, however, these questions relate to a fundamental need of the theory. We have already seen (in Chapter 3) the general kinds of reasons which prevent us from accepting countable additivity as an axiom. We shall come across other reasons which, taken with the above considerations, suggest that the use of Lebesgue measure and integration over the special field is not valid. (That is, of course, unless specific conditions are introduced in particular examples in order to meet the conditions which would allow such an application. The distinction, however, is that illustrated by the difference between saying 'I am applying this method because all functions are continuous', and 'I am applying this method because the function that I have chosen is continuous'.)

I do not know whether similar reservations and objections have a sound basis in regard to applications in other fields. In the case of mass, such a degree of detailed analysis is inappropriate (for instance, how could mass at rational points be separated off, or even considered as conceptually distinguishable?). The same thing could be said, in fact even more so, for statistical distributions. Everything leads us, therefore, to the conclusion that, apart from rather indirect issues,[10] the question is irrelevant in this area.

On the other hand, it is not really surprising if real objections only arise in the field of probability. In fact, we have, in the other fields, empirical assumptions, which are therefore approximate and necessarily lead to some arbitrariness in the mathematical idealization. The probabilistic interpretation, however, must confront logic face to face; this is its sole premise. Logic does not claim that it reaches out to some sort of precision (nor even to a higher level of approximation than is necessary), but neither does it allow the construction of a formally complete structure which does not respect the logical exigencies of a purely logical field of application; nor can it accept one constructed by someone else.

6.4 Distributions in Probability Theory

6.4.1. Let us now turn to the topic of direct interest to us: that is the application of these mathematical tools within the calculus of probability. Roughly speaking, the application takes the following form: for any random quantity X, one can imagine a distribution of probability over the x-axis by assigning to the distribution function the interpretation $F(x) = \mathbf{P}(X \leqslant x)$ = the probabilistic 'mass' on $[-\infty, x]$. It follows that $F(I) = \mathbf{P}(X \in I)$, and $F(\gamma) = \mathbf{P}(\gamma(X))$, for any set I and function γ for which the notation is applicable.

This formulation, which is deliberately rather vague and neutral, is intended as a curtain-raiser to the questions we shall have to consider later (perhaps it would be more accurate to say that we shall consider them in relation to our particular position). These basically concern the alternatives of either continuing with the Riemann framework, or abandoning it for that of Lebesgue. We shall, however, leave the way open for any further modifications that may be required.

10 Like that concerning the precise meaning of a differential equation expressing a physical law; or the definition of the integral on a contour having cusps (this topic has given rise to discussion about the Kutta and Joukowski theorem); and so on.

It is worth giving here and now a brief sketch of the two opposed positions. In order to pin a label on them, we might use the term *strong* for those who, as a result of accepting the validity of the Lebesgue procedures in this field, draw stricter and more sophisticated conclusions from the data; and *weak* for those who accept only the conclusions which derive from some smaller number of assumptions, carefully considered, and accepted only after due consideration.

The fact that we do not accept (as an axiom) countable additivity commits us to support of the *weak* position (as we have already mentioned, this is one of the main planks in our programme; see Chapter 1, 1.6.2–1.6.4). The present discussion, apart from giving more insight into the implications of not adopting countable additivity, will consider its relation to other topics, and, although confining itself to the simplest case of distributions on the real line, will, in fact, reveal the general import of the conclusions.

6.4.2. *The strong formulation*. Once we know $F(x)$ we know everything about the probability distribution of a random quantity X. Everything that can be defined in terms of $F(x)$ (and with the Lebesgue extension) has a meaning: nothing else does. The probability that $X \in I$ is either given by $F(I)$, if the set I is F-measurable (Lebesgue–Stieltjes), or has no meaning if I is not F-measurable. The same holds for the prevision of $\gamma(X)$: either the function $\gamma(x)$ is F-measurable, in which case $\mathbf{P}(\gamma(X)) = \boldsymbol{F}(\gamma)$, or the concept has no meaning. The set of possible values for X is also determined by F: it is the set of points for which F is increasing (i.e. the set of points not contained in an interval over which F is constant.[11])

In this approach, one operates entirely within the confines of a rigid formulation, prescribed in advance: it was to this type of structure that we applied the description 'Procrustean bed'. Within its confines, '*that which is not compulsory is forbidden*'.

6.4.3. *The weak formulation*. Knowledge of $F(x)$ is only one of the many possible forms of partial knowledge of the probability distribution of a random quantity X (although, in practice, it is one of the most important).

Complete knowledge would demand a 'complete distribution': in other words, a (finitely additive) extension of $F(\gamma)$ to every function γ (and, in particular, to every set I) with no restrictions (on integrability, measurability, or whatever) and such that

$$F(\gamma) = \mathbf{P}(\gamma(X))$$

always holds (in particular $F(I) = \mathbf{P}(\mathrm{X} \in \mathrm{I})$). Of course, we are talking of a theoretical abstraction, which can never actually be attained, but we have to make this the starting point, the landmark from which to get our bearings, in order to be in a position to consider all cases of partial knowledge without attributing to any of them some preordained special status.

Knowledge of $F(x)$, which we shall call *distributional* knowledge (or, sometimes, as is more common, knowledge of the distribution, albeit in the restrictive sense explained above), can turn out either to be more than we require, or less than we require, both

11 Even from this point of view, there would appear to be no difficulty in allowing something less rigid (e.g. the possibility of excluding a set of measure zero): I do not recall, however, ever having seen this kind of thing done explicitly. Perhaps this is the result of a psychological factor, which causes us to see distributions as prefabricated theoretical schemes, ready for attaching to random quantities, rather than regarding them as deriving from those random quantities, and from the particular circumstances which, depending on the case under consideration, derive from the underlying situation.

from the point of view of the possibility of determining it realistically, and in relation to the needs of the situation under study. Sometimes, $\mathbf{P}(X)$, or $\mathbf{P}(X)$ and $\mathbf{P}(X^2)$ together, or some other summary, may be sufficient; in such cases there is no need to look upon the distribution as the basic element from which all else follows. On other occasions, the distribution itself is not enough: this is the case whenever we wish to rid ourselves of the restrictions implicit in the properties of $F(x)$ as commonly accepted; restrictions which are not always appropriate.

In contrast to the strong formulation, the argument in the weak case is always developed with a great deal of freedom of action: there is no obligation to fill in more details of the picture than are strictly necessary, and, on the other hand, there is no limit to the extensions one can choose to make – even up to the (idealized) case of complete knowledge.

6.4.4. *Setting the discussion into motion*. We introduce straightaway some useful notation. Its present purpose is to enable us to distinguish between the various extensions we shall consider in relation to a given F; but it will also enable us to avoid repeated, detailed explanations, whose tendency (despite the intention of avoiding ambiguities) is rather to create confusion.

The general notation is as follows: if $\mathscr{G}$ is a given set of functions $\gamma(\gamma \in \mathscr{G})$, then F, thought of as defined on $\mathscr{G}$, will be denoted by $F_{\mathscr{G}}$; for every γ not in G, there will be for $F(\gamma)$ (used to denote a generic extension) a bound of the form $F^-_{\mathscr{G}}(\gamma) \leqslant F(\gamma) \leqslant F^+_{\mathscr{G}}(\gamma)$ (we do not dwell here upon the details of this : the interpretation is as set out in Chapter 3, 3.10.1 and 3. 10.7, and which will be of use to us later in 6.5.3). We shall adopt, for the time being, as special cases, the following notations for distinguishing the ambit over which F is thought of as defined:

$F_{\mathscr{R}}$: if relative to the Riemann field;
$F_{\mathscr{B}}$: if relative to the Lebesgue field;[12]
$F_{\mathscr{C}}$: if relative to the complete field; and, finally,
F : if used in a generic sense.

More precisely: $F_{\mathscr{B}}$ always denotes an F which has been extended to mean

$$F_{\mathscr{B}}(\gamma) = \int \gamma(x)\ \mathrm{d}F(x)$$

(in the Lebesgue–Stieltjes sense) where this makes sense; undefined otherwise. We could, however, denote the upper and lower integrals[†] by F^-_{B} and F^+_{B} and simply express the bounds $F^-_{\mathscr{B}}(\gamma) \leqslant F(\gamma) \leqslant F^+_{\mathscr{B}}(\gamma)$. In the above, $F_{\mathscr{B}}$ according to the *strong* formulation, *is all and everything:* the terms in the inequality do not even have a meaning within this framework. In the weak formulation, even if one considers an F which ('by chance', or for some particular reason – any reason – but not by virtue of some postulate) is countably additive over the Lebesgue field, the bounds would still have a meaning.

---

12 We shall use $\mathscr{B}$ instead of $\mathscr{L}$ (which was already used, see Chapter 2, for 'linear space'): $\mathscr{B}$, standing for *Borel*, is currently in use with a similar meaning to this (referring to Borel measure, which only differs from Lebesgue measure in so far as the latter extends it wherever it is uniquely defined by the two-sided bound). In our case $\mathscr{B}$ coincides with $\mathscr{L}$ (taken as meaning Lebesgue) because the extension is already implicit in our formulation ($F_{\mathscr{G}}(\gamma)$ is not only defined for $\gamma \in \mathscr{G}$, but for every γ such that $F^-_{\mathscr{G}}(\gamma) = F^+_{\mathscr{G}}(\gamma)$).

[†]For the time being, we are considering only those functions γ which are *bounded* (over the range on which F varies, namely the x for which $0 < F(x) < 1$). The other case will be dealt with specifically in Section 6.5.4.

$F_{\mathscr{C}}$ denotes any F whatsoever, finitely additive, and thought of as defined for all functions γ (the ideal case, thought of in the weak formulation as the basic landmark). In this case, it is clear that bounds on the indeterminacy do not make sense; neither is there any possibility of extension.

When it makes sense (when it does not we consider $F^{-}_{\mathscr{R}}$ and $F^{+}_{\mathscr{R}}$),

$$F_{\mathscr{R}}(\gamma) = \int \gamma(x)\, \mathrm{d}F(x),$$

in the Riemann–Stieltjes sense, expresses *all that one can obtain from F*; that is, *distributional knowledge*, according to the *weak* formulation:

$$F^{-}_{\mathscr{R}}(\gamma) \leqslant F(\gamma) \leqslant F^{+}_{\mathscr{R}}(\gamma).$$

We should make this more precise, but this first requires the following summary.

We summarize briefly the two opposing points of view which present (in terms of the notation introduced above) a choice between:

(*strong*): for a given X, an $F_{\mathscr{B}}$ is to be chosen, and there is nothing more to be said;
(*weak*): for a given X, an $F_{\mathscr{C}}$ should be chosen; in fact, one limits oneself to some partial $F_{\mathscr{G}}$ that serves the purpose; often, one chooses a distribution function $F(x)$, and then it follows that $F_{\mathscr{R}}$ is in $\mathscr{R}$, and that the bound, which lies between $F^{-}_{\mathscr{R}}$ and $F^{+}_{\mathscr{R}}$ is not in $\mathscr{R}$.

6.4.5. Once more a word of warning. When referring to distributions, or distribution functions, F, it is useful to think of them as mathematical entities (e.g. the function $F(x) = ½ + (1/\pi)$ arctan x), which are available for representing the probability distribution of any random quantity X, as required. In other words, it is better *not* to think of them as associated with any given X. This distinction is of a psychological nature rather than a point of substance – which explains why the explanation is vague and somewhat confused – but our aim is to warn against misunderstandings that can (and frequently do) arise through some sort of 'identification' of an $F(x)$, an abstract entity, with $P(X \leqslant x)$, which, although equal to it, is a concept dependent on the specific random quantity X that figures in it. A typical example of the misunderstandings to be avoided is the confusion between limit properties of a sequence of distributions and similar behaviour of random quantities which could be associated with those distributions.

6.4.6. Why the 'Procrustean bed'? A preliminary question which it might be useful to discuss (although more for conceptual orientation than as a real question) is the following. Why is it that, at times, some people prefer (as in the *strong* formulation) to adopt a fixed frame of reference, within which one assumes complete knowledge of everything, all the details, no matter how complicated, no matter how delicate, and irrespective of whether they are relevant or not? This, despite the fact that the system is only used to draw particular conclusions, which could have been much more easily obtained by a direct evaluation. All this would appear to be a purely academic exercise; far removed from realism or common sense.

In seeking the reason for this, one should probably go back to the time when fear was the order of the day, and all manner of paradoxes and doubts resulted. The only hope of salvation was to take refuge within paradox-proof structures – and this was no doubt right, at the time.

We must consider, however, whether it is reasonable, or sensible, to force those who are now strolling across a quiet park to take the same precautions as the pioneers who

originally explored the area when it was wild and overgrown, and were ever fearful of poisonous snakes in the grass?

Let us note the following in connection with a specific example:

> the use of transfinite induction (Chapter 3, 3.10.7) assures us that we can always proceed in an 'open-ended' way, adding in new events and random entities from outside any prefabricated scheme;
>
> this method of proceeding is the only sensible one; at any moment new problems arise, and the thought of someone having to unscramble the enormous Boolean algebra that he has fixed in his mind, together with the probabilities which are stuck on all over the place, and having to construct a new edifice in order to include each new event, each new piece of information, and to update all his probabilities before sticking them back in, this thought is horrifying;
>
> in evaluating probabilities (or a probability distribution), one should also proceed step by step, making them, little by little, more and more precise, for as long as it seems worth continuing. Even Ovid did not record the sudden appearance of a complete Boolean algebra, armed with all its probabilities, and springing from the head of Jove, disguised as Minerva, or rising, like Venus, from the foaming sea.

These remarks have been expressed in a manner which accords with the subjectivistic point of view; they would seem, however, to reflect fully the requirements of any realistic point of view, although perhaps not in such a clear-cut manner.

6.4.7. *The absence of anything having a special status.* We have already said (in 6.4.3) that no partial knowledge was to be accorded special status: not even that provided by $F(x)$. It seems strange to deny special status to probabilities associated with the 'most basic' sets, like intervals (or with continuous functions, as opposed to sets or functions of a 'pathological' nature). Is this objection well founded? Nothing can really be said about this without first considering and analysing the sense in which something has to be 'true', and in what sense, and on what basis, things appear to us as strange or pathological.

With regard to our own enquiry, we must distinguish that which has a *logical* character from that which draws its meaning from *other* sources; this is necessary, because it is only differences of a logical nature which can lead to the possibility of different treatment from a logical point of view. We note, therefore, that, from a logical point of view, in this representation every event corresponds to a set of points, and the only property that is relevant is the fact that one can tell (on the basis of the occurrence of X) whether the 'true' point belongs to the set or not. In this sense, there is nothing that can give rise to special forms of treatment: the above-mentioned property is assumed to be valid everywhere by definition, and other properties do not enter into consideration. From a logical point of view, no other aspects are relevant; for example, topological structures, or some other kind of structures that the space may happen to have for reasons which do not concern us.

Only differences of a logical nature could possibly justify special treatment in a probabilistic context. In general, there is no reason to discriminate between sets, and, in particular, this applies to sets which have, with respect to the outcomes of a random quantity X, the form of intervals, or anything else, however 'pathological'. There is no justification for thinking that some events merit the attributing of a probability to them, and others do not; or that over some particular partitions into events countable additivity holds, but not over others; and so on.

6.4.8. *The argument concerning what happens 'outside $\mathscr{B}$'.* We know that countable additivity cannot hold over the entire field $\mathscr{C}$ (of all events $X \in I$ and random quantities $\gamma(X)$ which can be defined in terms of a random quantity X, in correspondence with all sets I and functions γ). In fact, this was proved by Vitali under the additional assumption of invariance for the measures of *superposable* sets; an assumption which was removed in the extensions mentioned previously.

The above could be taken in itself as a sufficient reason *for rejecting countable additivity as a methodologically absurd condition* (as a general, axiomatic kind of property) since it sets itself against the absence of any logical distinctions, which alone could justify discrimination between events.[13] This would be the case even if we disregarded the reasons we have already put forward (Chapter 3, 3.11, and Chapter 4, 4.18), reasons which, in fact, cannot be disregarded.

6.4.9. *The argument about what happens 'inside $\mathscr{B}$'.* In the particular case of L-measurable sets, where we know that countable additivity *can* be assumed without giving rise to any contradictions, there is no reason to assume automatically that countable additivity *must* hold (or that it is entitled to be accepted for some particular reason). Every distinction between measurable and nonmeasurable sets disappears when we no longer take the topology of the real line into account (imagine reshuffling the points as though they were grains of sand). We present straightaway some counterexamples (they can be disposed of only on the grounds of a prejudice to do so just because they are counterexamples[14]).

Here is one of them. Let X be a rational number between 0 and 1, and let us further assume that no rational between 0 and 1 can either, on the basis of our present knowledge, be rejected as impossible, or appear sufficiently probable to merit assigning a nonzero probability to it. In this case, we have a continuous distribution function $F(x)$: we could also limit ourselves to considering the special case of the uniform distribution, $F(x) = x$ $(0 \leqslant x \leqslant 1)$. According to the strong formulation, one would conclude that, with probability 1, the rational number X belongs … to the set of irrational numbers!

This, and other similar examples (which we shall make use of shortly for other purposes), also show, among other things, that precisely the same distribution function can correspond to random quantities having different ranges of possible values. This will be dealt with in Sections 6.5.2–6.5.3.

6.4.10. *Partial knowledge.* Every piece of partial knowledge will be the knowledge of the complete distribution $F_{\mathscr{C}}(\gamma)$ restricted to some subset or other of the functions γ (it does not matter whether they are functions, sets, or a mixture of the two). For example, one might know $F(x)$ at some particular points (i.e. for a certain partition into intervals) and/or γ for some individual functions. To use the standard examples, these might

---

13 More precisely, the discrimination would only be justified if one concentrated the whole probability (=1) on a finite or countable set (of points with positive probabilities. with sum 1). It is absurdly restrictive to pretend this should always be the case; even, e.g., if the 'points' of our field are 'all the possible histories of the universe' (but let us leave aside such extralogical and personal judgements). The fact is, that no continuous measure – in the mild sense of being, like Lebesgue measure, effectively spread over an uncountable set – can satisfy our requirement.

14 This is the tactic of 'monsterbarring', according to the terminology of Imre Lakatos, in "Proofs and refutations", *Brit. J. Philosophy of Science*, **14** (1963–64), 53–56.

be the prevision and variance (as direct data, and not based on the assumption, either implicit or explicit, of the existence of the distribution of which the prevision is the barycentre etc., as is usually the case). It would, however, be equally admissible (although, generally speaking, of little interest, and not really practicable) to provide, instead, probabilities for certain pathological sets only (e.g. numbers whose decimal expansions never involve more than n zeroes in the first $2n$ places), or the previsions of some pathological functions (e.g. continuing with the same example, $\gamma(x)$ = sup of the percentage of zeroes in the first n places as n varies).

In short, it is open to us to assume or require that either *everything*, *a little* or *a great deal* is known about the probabilities and previsions relating to X. Do not lose sight of the fact (even though it is not convenient to repeat it too frequently) that, in using 'known' or 'not known' when thinking in terms of the mathematical formulation (in fact, when thinking of the actual meaning), we mean 'evaluated' or 'not evaluated'.

Of course, it could be, as a special case, that the partial knowledge of the complete distribution is that defined over the intervals: in other words, that given by $F(x)$, known for all x. This is what we have called knowledge of the distribution through the *distribution function*. It is a form of partial knowledge like all the others but it is of particular interest and we shall wish to, and have to, consider it at greater length, in order to clarify the rôle played (in the present formulation) by $F(x)$.

$F(x)$ remains a standard tool, but re-evaluated (one might say cut down to size) in a manner and for reasons that we shall explain. It does not play any special, privileged rôle *de jure*, but only *de facto:* that is, in relation to the interpretation of X as a magnitude, which is what is of interest in practice, and to the geometric representation on the line, which is what enables it to be visualized. It is for these reasons that it plays a special rôle, by reason of the applications, and from the psychological point of view; despite the fact that they cannot justify its special *status* from the logical standpoint.

6.4.11. The *re-evaluation* is not solely, however, in this conceptual specification; nor in the fact that knowledge conveyed by $F(x)$ no longer appears complete in that we require something further ($F(\gamma)$ lying outside the Lebesgue ambit of F), whereas it remains what it is. But it does not remain what it was: it is more restricted. It remains what it was only in the Riemann ambit of F; outside of this (with no further discrimination between that which is inside or outside the Lebesgue ambit of F) it only provides the bounds we have already encountered

$$F^{-}_{\mathscr{R}}(\gamma) \leqslant F(\gamma) \leqslant F^{+}_{\mathscr{R}}(\gamma).$$

These give the limits for any evaluation of $F(\gamma)$ compatible with knowledge *of* F in the distributional sense (i.e. knowledge of $F(x)$). We are, of course, dealing with the upper and lower integrals in the Riemann sense; in particular (in the case of sets) we have inner and outer Jordan–Peano measure. This indeterminacy does not imply any fault in the capacity of the concepts to produce a unique answer; on the contrary, as we shall see later in more detail, the indeterminacy turns out to be essential (given our assumptions), in the sense that all and only the values of the interval are in fact admissible (and all equally so). Any of them can be chosen, either by direct evaluation, or by an evaluation which derives from some additional considerations, which must then be set out one by one (and cannot just consist of the assumption of countable additivity, for which one must, case by case, make the choice of the family of partitions on which its validity is to be assumed, and state the choice explicitly).

What we have said so far concerning the rôle of $F(x)$ is more or less the translation and explication in concrete form of the two 'reservations' that we previously put forward in the abstract. But the abandonment of countable additivity implies yet another revision of the meaning of $F(x)$: it is no longer true that a jump at x must correspond to a concentration of probability at the point x (it may only *adhere* to the point, and the point itself might not even belong to the set of possible points). It is also no longer true that $F(x)$ must vary from 0 to 1 (we only require that $0 \leqslant F(-\infty) \leqslant F(+\infty) \leqslant 1$), or that the possible points are those at which $F(x)$ is increasing.

A single observation will suffice. Suppose that the possible points, judged equally likely, form a sequence (e.g. $x_0 - 1, x_0 - \frac{1}{2}, \ldots, x_0 - 1/n, \ldots$) which tends to a given point x_0 from below. In this case $F(x)$ will have a jump of 1 at $x = x_0$, just as if $X = x_0$ with certainty (all the mass concentrated at x_0). In fact, we have $F(x) = (x \geqslant x_0) = 0$ for $x < x_0$, and = 1 for $x \geqslant x_0$, because to the left of any point on the left of x_0 there is at most a finite number of possible points, each of which has zero probability; whereas to the left of x_0 (and, *a fortiori*, to the left of any point on the right of x_0) we find all the possible points.

This implies that, in general, if $F(x)$ has a jump p_h at a point x_h, it is always possible (apart from the case when there are no possible points in some left or right neighbourhood of x_h) to decompose p_h, in some way, in the form $p_h = p_h^- + p_h^0 + p_h^+$, where p_h^0 is the mass actually concentrated at x_0, and the other two parts are *adherent* to it on the left and on the right (in the manner illustrated in the example).

This fact alone would seem to provide support for the usefulness of the convention of regarding the value of $F(x)$ to be indeterminate at points of discontinuity (see footnote 3). We shall, however, consider this in the next section (6.5.1), where the arguments will be more decisive when put in the context of some further ideas.

The previous example (if we consider sequences tending to $-\infty$ or to $+\infty$) suffices to show that we can, in a similar fashion, have probabilities adherent to $-\infty$ and to $+\infty$. These are given by $F(-\infty)$ and $1 - F(+\infty)$. Those distributions for which (as we have so far assumed, in accordance with the standard formulations) these probabilities are zero we shall call *proper*, and we note that F then actually does vary between 0 and 1; all others will be called *improper* (and we can further specify whether the impropriety is *from below from above* or *two-sided*).

Our previous remark concerning possible points is also clear, given the possibility of substituting for any point a sequence which converges to it; this topic will be considered further in due course (see 6.5.2).

6.5 An Equivalent Formulation

6.5.1. Knowledge of $F(x)$ (apart from points of discontinuity), in other words, what we are calling distributional knowledge, is equivalent – in the case of a *proper* F[15] – to knowledge of $F(\gamma)$ for all continuous functions γ, which are bounded over the entire x-axis, from $-\infty$ to $+\infty$. More precisely, these, and only these, functions are F-integrable whatever F might be; conversely, knowledge of $F(\gamma)$ for all continuous γ is sufficient, whatever F might be, to determine $F(x)$ for all x, apart from discontinuity points.

15 Otherwise one requires in addition the existence of a finite limit for $\gamma(x)$ as $x \to -\infty$, or $x \to +\infty$, or both.

Of course, to say that knowledge of $F(x)$ is equivalent to knowledge of $F(\gamma)$ for all continuous γ does not mean that it has to be known for every such γ. It will be sufficient to know it for a basis in terms of whose linear combinations any continuous function can be approximated. This remark will serve as the foundation for more analytical kinds of treatment (in particular, that for characteristic functions); here it merely serves to assuage possible doubts.

Let us consider the following in more detail, further considering the possibility of 'adherent masses', which we noted above. If $F(x)$ has a jump p_h at the point $x = x_h$, and it were assumed that the mass p_h were concentrated at the point x_h, then (as in the case of the usual assumption of countable additivity) we would take the contribution of this mass to $F(\gamma)$ to be $p_h\gamma(x_h)$. Without the assumption of concentration, however, we can do no more than note that the contribution lies between the maximum and minimum of the five values

$$p_h\gamma(x_h) \quad \text{and} \quad \left.\begin{matrix}\max\\ \min\end{matrix}\right\} \lim p_h\gamma(x)$$

as $x \to x_h$ from the left or right, respectively. Proceeding differently (and more simply) it is sufficient to exclude points of discontinuity as subdivision points (this is always possible – there are only a countable number of them).

From this, it is clear that any function $\gamma(x)$ that has even a single discontinuity point is not integrable for all F, since, if we take an F with a jump at this point, the contribution of this mass to the integral is indeterminate. Conversely, if we know $F(\gamma)$ for the continuous functions γ, we can evaluate $F(x_0)$ from below and above as follows: we take a function $\gamma_1(x)$ which = 1 from $-\infty$ to $x_0 - \varepsilon$, and = 0 from x_0 to $+\infty$, and decreases continuously from 1 to 0 within the small interval $x_0 - \varepsilon$ to x_0, and a function $\gamma_2(x) = \gamma_1(x - \varepsilon)$, which is the same as γ_1, except that the decreasing portion is now between x_0 and $x_0 \pm \varepsilon$. The difference between the two functions is $\leqslant 1$ between $x_0 \pm \varepsilon$ and zero elsewhere; we therefore have that

$$F(\gamma_2) - F(\gamma_1) \leqslant F(x_0 + \varepsilon) - F(x_0 - \varepsilon), \text{ etc.}$$

Everything goes through smoothly, except when we have a discontinuity at $x = x_0$.

The mathematical argument, which seems to me to show conclusively that we should consider $F(x)$ as indeterminate at discontinuity points x, is the following: it is more meaningful to consider the continuous γ, than to consider indicator functions of half-lines or intervals. What seemed to be an ad hoc restriction when starting from the intervals, is, instead, rather natural when one considers continuous functions; in this case, one would need an ad hoc convention to eliminate it.

On the other hand, this mathematical argument is closely bound up with the point that I consider to be most persuasive both from the point of view of fundamental issues and of applications: the need for some degree of realism when we assume the impossibility of measuring X with absolute certainty. We shall consider in the Appendix (Section 7) limitations imposed on 'possible occurrences' of events due to these kinds of imprecisions; it is clear, however (and we shall confine ourselves to this one observation at present), that to consider $F(x)$ as completely determined, apart from discontinuity points, is equivalent to thinking that X can be measured with as small an error as is desired, but cannot be measured exactly with error = 0. This suffices to render the case

$X = x_0$ with certainty indistinguishable from the case where the mass is adherent to x_0 (e.g. it is certainly at $x_0 - 1/n$, where n is any positive integer whose probability of being less than any preassigned N is equal to zero).[16]

6.5.2. *The distribution and the possible points.* We have already seen, when examining the special case of a discontinuity point, that there is a lot of arbitrariness concerning the possible points which 'carry' the mass corresponding to the jump; they do not have to enclose the jump-point, they only have to be dense in any neighbourhood of it. Before proceeding any further, we have to examine the general relationship between the set $\mathscr{P}$ of possible points for a random quantity X – which we shall call the *logical support* of X – and $F(x)$, the distribution function of X; more specifically, the relationship between this set $\mathscr{P}$ and the set $\mathscr{D}$ of points at which $F(x)$ is increasing – which we shall call the *support of the distribution* F (or the *distributional support* of X). Formally, this is the set of x such that, for any $\varepsilon > 0$, we have

$$F(x+\varepsilon) - F(x-\varepsilon) > 0.$$

Every neighbourhood of x has positive probability; it is therefore possible, and hence contains possible points. It therefore follows that $\mathscr{D}$ is contained in the closure of $\mathscr{P}$; moreover, this condition is sufficient because, whatever partition one considers (partition, that is, of the line into intervals), no contradiction is possible (every interval with positive mass contains possible points to which it can be attributed).

It is convenient to consider separately the various cases. Let us begin with the intervals on which $F(x)$ is constant (at most a countable collection): these may contain no possible points but there is nothing that debars them from doing so (they could consist entirely of possible points), so long as the total probability attributed to them is zero. At the other extreme, we have the intervals over which $F(x)$ is strictly increasing. Here, it is necessary and sufficient that the possible points are everywhere dense (it could be that all points are possible). As an example, think of the uniform distribution on [0, 1], with either all points possible, or just the rationals. An isolated point of increase is necessarily a jump-point (but not vice versa), and we have already discussed this case; either the point itself must be possible, or there must exist an infinite number of possible points adherent to it (of which it is a limit point). Finally, suppose that a point of increase of $F(x)$ is such because each neighbourhood of it contains intervals, or isolated jump-points, where $F(x)$ is increasing. This fact tells us that the given point is an accumulation point of possible points; we can go no further in this case.

We are especially interested in the end-points of the above-mentioned sets. We have adopted (ever since Chapter 3) the notation inf X and sup X for the limits of the logical support; let us now denote by inf F and sup F the limits of the distributional support. These are, respectively, the maximum value of x such that $F(x) = 0$, and minimum value

16 Without going into the theoretical justifications (or attempts at justifications), it is a fact that the different conventions reveal practical drawbacks that make their adoption inadvisable. The convention $F(x) = F(x + 0)$ (or, conversely, $F(x) = F(x - 0)$) makes the equation $F_1(x) = 1 - F(x)$ (used in passing from X to $-X$) invalid; writing $F(x) = \frac{1}{2}[F(x+0) + F(x-0)]$ avoids this difficulty, but (see the end of 6.9.6) one sometimes needs to consider $F_2(x) = [F(x)]^2$, and it is not true that $\{\frac{1}{2}[F(x+0) + F(x-0)]\}^2 = \frac{1}{2}[F^2(x+0) + F^2(x-0)]$; and so on. In contrast, the convention we are proposing here remains coherent within itself; moreover, it gives a straightforward interpretation of the appropriateness of completing the diagram of Figure 6.2a (Figure 6.2b) with vertical (horizontal) segments.

such that $F(x) = 1$ (if F is unbounded – from below, from above, or from both sides – or improper, the values are $\pm\infty$). By virtue of what we said previously, we necessarily have $\inf X \leqslant \inf F \leqslant \sup F \leqslant \sup X$. It is important to note that logical support is a bound for distributional support, but not conversely.

More generally, it is important to realize just how weak the relation between the two forms of support can be. If we are given the distribution, all we can say is that each point of the support is either a possible point, or is arbitrarily close to possible points; in addition to this, possible points (with total probability zero) could exist anywhere and even fill up the whole real line. On the other hand, given the logical support, we can state that the distribution could be anything, so long as it remains constant over intervals not containing any possible points. We are here merely reiterating, in an informal and rather imprecise way, what we have already stated precisely. In this way, however, we may be able to better uncover the intuition lying behind the conclusions. On the one hand, that, corresponding to the concept of being able to take measurements as precisely as one wishes, but not exactly, one is indifferent to the fact that what is regarded as possible can be: either a point or a set of points arbitrarily close to it, respectively; either all the points of an interval or those of a set everywhere dense in it, respectively. On the other hand, that possible points with total probability zero do not affect the distribution, but are not considered as having no importance (and we shall see below that they are important when it comes to considering prevision).

6.5.3. Conclusions reached about sets lead immediately to conclusions regarding their probabilities. In fact, we can see straightaway that $\mathbf{P}(X \in I)$, the probability of a set I, can actually assume any value lying between the inner and outer F-measure (in the Jordan–Peano sense).

Let $\mathscr{D}$ be the set of points for which $F(x)$ is increasing, and partition it into $\mathscr{D}_1$, the intersection of $\mathscr{D}$ with the closure of I (that is, the set of points of $\mathscr{D}$ having points of I in every neighbourhood), and $\mathscr{D}_2$, its complement (points within intervals containing no points of I). Let us assume that in the closure of $\mathscr{D}_1$ only points of I are possible (either all of them, or a subset which is everywhere dense there); only in the intervals containing no points of I do we have recourse to other points in order to obtain the 'possible points' required for $\mathscr{D}_2$. In this way, I turns out to have the maximum possible probability; that is, the outer F-measure (we attribute to I the measure of every interval in which I is dense). By applying the same idea to the complement of I, we obtain the other extreme (the minimum probability for I, given by the inner F-measure; in this case only those intervals containing solely points of I are considered). Clearly, all intermediate cases can be arrived at by mixtures (for example, for a direct interpretation, consider the fact that, without changing the distribution, possible points are taken either to be those of the first version or the second, depending on whether an event E is true or false; by varying the value $p = \mathbf{P}(E)$, $0 \leqslant p \leqslant 1$, we obtain all possible mixtures).

This fact reveals another aspect of the 're-evaluation' of the nature of distributional knowledge: it says very little about what, from a logical viewpoint, is the most important global feature of the distribution; that is, about the logical support.

6.5.4. *The restriction of boundedness.* There remains the question of our restriction to the bounded case: it is an important topic in its own right and we have rather passed it over (each topic should really come before all the others, and that is just not on). We shall meet a further aspect (the last one!) of the 're-evaluation' of the role of the distribution

function and we shall be forced to make (and offer to the reader) some sort of makeshift choice, not entirely satisfactory, in order to be able to draw attention to certain necessary distinctions, without too many annoying notational complications, and without running too many risks of ambiguity.

We have already seen (Chapter 3, 3.12.4–3.12.5) that, without the assumption of countable additivity, there are no upper (lower) bounds for the prevision of a random quantity which is unbounded from above (below). This was seen in the case of discrete random quantities; what happens when we pass from this to the general case?

The question is an extremely deceptive one when looked at in the light of what distributional knowledge is able to tell us. Starting from the knowledge of $F(x)$, the conclusion that we can derive a certain value, $F(\Box)$, which 'ought to be' that of $\mathbf{P}(X)$, will be more acceptable if not only the distribution F, but also the logical support of X, is bounded (and knowledge *of* F gives us no information about this). We shall put this conclusion more precisely, and also examine more closely the value of the partial knowledge that we can obtain in this connection.

First of all, it is convenient to specialize to the case of non-negative random quantities ($\inf X \geqslant 0$): given any X, we can, of course, decompose it into the difference of two non-negative random quantities by setting

$$X = X\big(X \geqslant 0\big) + X\big(X \leqslant 0\big),$$

or, in a different but equivalent form,

$$X = \big(0 \vee X\big) + \big(0 \wedge X\big).$$

In either case, the first summand has value X if $X \geqslant 0$, and zero otherwise; and the second summand has value X if $X \leqslant 0$ and zero otherwise (and is therefore always nonpositive: in order to obtain the difference of nonnegative values explicitly, it suffices to write 1st – (–2nd) instead of 1st + 2nd).

For X non-negative and bounded, we certainly have

$$\mathbf{P}\big(X\big) = F\big(\Box\big) = \int x \ \mathrm{d}F\big(x\big).$$

A non-negative X that is unbounded can be turned into a bounded quantity by either 'amputating' or 'truncating' it.[17] We shall apply the first method, which is simpler. We have

$$\mathbf{P}\big(X\big) \geqslant \mathbf{P}\Big[X\big(X \leqslant K\big)\Big] = F\Big[\Box\big(\Box \leqslant K\big)\Big] = \int_0^K x \ \mathrm{d}F\big(x\big);$$

this holds for any K, and hence

$$\mathbf{P}\big(X\big) \geqslant \int_0^\infty x \ \mathrm{d}F\big(x\big) = F\big(\Box\big),$$

where this defines $F(\Box)$ by convention in this case. The integral may be either convergent or divergent: in the latter case, we must have $\mathbf{P}(X) = F(\Box) = +\infty$, whereas, in the

17 To 'amputate' means to put $Y = X(X \leqslant K)$: to 'truncate' means to set $Z = X \wedge K$; in other words, $Y = Z = X$, so long as $X \leqslant K$, but $Y = 0$ and $Z = K$ otherwise. Clearly we have $Y \leqslant Z \leqslant X$ ($Y = Z = X$ if $X \leqslant K$, and $Y < Z < X$ when $X > K$, since $0 < K < X$).

former, we can only say that all values in the range $F(\square)$ to $+\infty$ are possible for $\mathbf{P}(X)$ (including the two extremes). Note that the case of convergence also includes the case where the distribution is bounded ($\sup F < \infty$), but arbitrarily large possible values of X (with total probability 0) are permitted.

6.5.5. We have adopted as a *convention* the definition $F(\square) = \int x \, \mathrm{d}F(x)$; this holds even when the integral is improper (it has to be extended up to $+\infty$) and only makes sense, as a limit, when it converges. This convention can be extended to the general case (to a distribution unbounded either way) with a similar interpretation; that is, with the understanding that

$$\int = \int_{-\infty}^{0} + \int_{0}^{+\infty},$$

if both integrals exist. We have to stress the interpretation we give to our convention, in order to draw a distinction between it and the interpretation it has in the usual formulation (that is, in the strong formulation). In the latter, the convention is taken as *a definition of the prevision* $\mathbf{P}(X)$ *of a random quantity X with distribution* $F(X)$*:* if one of the two integrals diverges, we either have $\mathbf{P}(X) = \infty$ or $\mathbf{P}(X) = -\infty$; if both diverge, $\mathbf{P}(X)$ has no meaning.

So far as we are concerned, $\mathbf{P}(X)$ will from henceforward have the meaning we have assigned to it; it will not make sense to set up new conventions in order to redefine it for this or that special case. Given the knowledge of $F(x)$ one could work out possible bounds for $\mathbf{P}(X)$ – always on the basis of the (weak) conditions of coherence – but one must be careful not to add any further restrictions and not to interpret the acceptable ones as being in any way more restrictive than they actually are. Not a single one of the values that can be attributed to $\mathbf{P}(X)$ without violating coherence should be ruled out as unacceptable. This would be a mistake; excusable if due to an oversight, but inexcusable if due to carelessness, or an inability to understand the demands of logical rigour.

Our convention should be interpreted entirely differently. It defines $F(\square)$ – and, similarly, $F(\gamma)$, for any γ – as information relating to the distribution F (considered as a mathematical entity); in order to avoid any confusion, it would perhaps be better to call $F(\square)$ the *mean value* of the *distribution F*, rather than the *prevision* (a notion concerning a random quantity X). Such a *mean value* is of interest when we are considering the previsions of random quantities X, Y, Z, all having the same distribution F; it is almost never possible, however, to simply state that the previsions must all be equal and coincide with $F(\square)$.

This conventional mean value does, however, play an important role for the following three reasons. In the first place, it serves to provide the logical conditions that characterize the set of admissible values for $\mathbf{P}(X)$. Secondly, it always provides a particular admissible evaluation of $\mathbf{P}(X)$, whose acceptance can often be justified by making an additional, meaningful assumption. Thirdly, it turns out that simultaneously accepting this additional assumption for several random quantities cannot lead one into incoherence.

If there is no additional knowledge, there are no logical conclusions to be drawn in passing from $F(x)$ to $\mathbf{P}(X)$. Fortunately, knowledge is available concerning a basic fact of a logical nature: that of the logical support of X (the set of possible values), or simply knowledge of the extremes, $\inf X$ and $\sup X$, or, even more simply, knowledge of whether they are finite or infinite. If they are both infinite, nothing more can be said about

$\mathbf{P}(X)$ – all values $-\infty \leqslant \mathbf{P}(X) \leqslant +\infty$ are admissible. If they are both finite, we must certainly have $\mathbf{P}(X) = F(\square)$. If only one of the extremes is infinite, all values between it and $F(\square)$ are admissible; in other words, if $\inf X = -\infty$, we have $-\infty \leqslant \mathbf{P}(X) \leqslant F(\square)$, and, if $\sup X = +\infty$, we have $F(\square) \leqslant \mathbf{P}(X) \leqslant +\infty$. In just one special case we also have a uniquely determined value: if $F(\square) = +\infty$ and $\inf X > -\infty$, then we certainly have $\mathbf{P}(X) = +\infty$ (and, similarly, if $F(\square) = -\infty$, and $\sup X < +\infty$, then $\mathbf{P}(X) = -\infty$).

Turning to the case of arbitrary functions, $\gamma(x)$, there are no essential changes to be made, but there are a couple of details.

In order to remain within the domain of distributional knowledge, we must limit ourselves to considering $F_{\mathscr{R}}$ (integrals in the Riemann–Stieltjes sense etc.) and, hence, to consideration of γ which are continuous (see 6.5.1), or, alternatively, to considering the two values $F^{-}_{\mathscr{R}}(\gamma) \leqslant F^{+}_{\mathscr{R}}(\gamma)$ (which are, in general, different). We shall always adopt the latter course, and, consequently, we will omit the $\mathscr{R}$. Extension to unbounded $\gamma(x)$ has to proceed as above; by separating into positive and negative parts, $\gamma(x) = [0 \vee \gamma(x)] + [0 \wedge \gamma(x)]$, and then amputating each of the parts (considering, for example, $[0 \vee \gamma(x)]\cdot[\gamma(x) \leqslant K]$ instead of $0 \vee \gamma(x)$; we shall call this $\gamma_K(x)$): we then take $F^{-}(\gamma_K)$ and $F^{+}(\gamma_K)$ relative to these, and obtain $F^{-}(0 \vee \gamma)$ and $F^{+}(0 \vee \gamma)$ as limits as $K \to \infty$. Similarly, we deal with $0 \wedge \gamma$, taking $K < 0$ and tending to $-\infty$. Summing, we obtain $F^{-}(\gamma) = F^{-}(0 \vee \gamma) + F^{-}(0 \wedge \gamma)$ (and similarly for F^{+}). If the sum is of the form $\infty - \infty$, it must obviously be understood as $-\infty$ for $F^{-}(\gamma)$ and $+\infty$ for $F^{+}(\gamma)$.

The second detail (perhaps it would be better to call it a remark) concerns a simplification that can arise in the case of an arbitrary $\gamma(x)$, in comparison with the simplest case, $\gamma(x) = \square(x) = x$, considered above. In fact, if the function y is bounded ($|\gamma(x)| \leqslant K$ for all x) then $\gamma(x)$ is certainly also bounded (and the same holds for semi-boundedness). If $\gamma(x)$ is not bounded, and all the values of x ($-\infty \leqslant x \leqslant +\infty$) are possible for X, then the random quantity $\gamma(X)$ is also unbounded, in the same manner. It is only in the case of $\gamma(X)$ unbounded and X having a more restricted support that the question of the boundedness of $\gamma(X)$ cannot be settled immediately, but only by examining the values that $\gamma(X)$ assumes on the support of X (it will often, however, be sufficient to check whether it is bounded on the interval $\inf X \leqslant x \leqslant \sup X$; only if it does not turn out to be bounded there will it be necessary to proceed to a more detailed analysis).

6.5.6. This having been said, our previous conclusions, apart from obvious changes, can now be restated, a little more concisely, in the general case.

The admissible values for $\mathbf{P}[\gamma(x)]$ *are those which satisfy the inequality*

$$F^{-}(\gamma) \leqslant \mathbf{P}\left[\gamma(X)\right] \leqslant F^{+}(\gamma)$$

when $\gamma(X)$ *is bounded* (*that is, if* $-\infty < \inf \gamma(X), \sup \gamma(X) < +\infty$); *with* $F^{-}(\gamma)$ *replaced by* $-\infty$ if $\inf \gamma(X) = -\infty$; *with* $F^{+}(\gamma)$ *replaced by* $+\infty$ *if* $\sup \gamma(X) = +\infty$.

In other words: in the double inequality, the right-hand side, left-hand side, or both, must be suppressed according to whether we have unboundedness on the left, right or both.

In particular, we obtain a uniquely determined value for $\mathbf{P}(X)$ only if $F(\gamma)$ exists (that is, $F^{-}(\gamma) = F^{+}(\gamma)$). This value *is finite* if $\gamma(X)$ is bounded; *infinite* ($-\infty$ or $+\infty$) if $\gamma(X)$ is semi-bounded (the direction of the boundedness is obvious).

To see how the present statement contains the previous one as a special case, observe that if both the integrals (from $-\infty$ to 0 and from 0 to $+\infty$) diverge, then $F^{-}(\square) = -\infty$ and $F^{+}(\square) = +\infty$.

6.5.7. *Prevision viewed asymptotically*. If $F(x) = \mathbf{P}(X \leqslant x)$, the mean value of the distribution F, in addition to its logical interpretation within the confines discussed above, may often have a reasonable claim to be taken as the value of $\mathbf{P}(X)$, even if there are no circumstances compelling one to make this choice.

This is the case when we choose to deal with an unbounded distribution (either one-sided or two-sided), but where the choice might reasonably be seen as an idealized approach to something that, had we been more realistic, should be considered as bounded. To put it more straightforwardly: we think that $F(x)$ represents pretty well our idea of the distribution throughout the range $a \leqslant x \leqslant b$, which, practically speaking, includes all the possible values; to also include the 'tail' to infinity is both convenient from a mathematical point of view, and also in practice, since we would not really know just where to set the limits a and b (but this latter point should not be taken too seriously). The most appropriate 'model' is to conceive of using the bounded distribution as 'a limit case of distributions amputated or truncated to intervals, whose limits are so large that an asymptotic expression is appropriate' (that is, for $a \to -\infty$ and $b \to +\infty$, in whatever way).

From among the logically admissible values for $\mathbf{P}(X)$ we shall often select this one when such justifications of asymptotic kind appear to be valid. Sometimes we shall denote this value by $\hat{\mathbf{P}}(X)$: the accent will simply signify that this particular choice has been made (it serves as a shorthand) and will not imply that $\mathbf{P}$ has been thus marked because it is a special value of some sort.

We have stated already that there is no danger of contradiction resulting from the systematic use of $\hat{\mathbf{P}}$; this means that $\hat{\mathbf{P}}$ is additive.

(We observe that in choosing values for $\mathbf{P}(X)$, $\mathbf{P}(Y)$ and $\mathbf{P}(Z)$, it is not enough merely to ensure that each of them is admissible – for example, if we have $Z = X + Y$ with certainty, then our choice must satisfy $\mathbf{P}(Z) = \mathbf{P}(X) + \mathbf{P}(Y)$.)

That this condition is satisfied for $\hat{\mathbf{P}}$ follows from the additivity of the integral. We are, however, dealing with a two-dimensional distribution, and we shall therefore deal with this later (in Sections 6.9.1–6.9.2).

In order to avoid unnecessary complications, we shall, unless otherwise stated, adopt the convention that we shall always take $\mathbf{P} = \hat{\mathbf{P}}$ (exceptions will be made when there is some critical remark worth making). Important points will be made in Section 6.10.3, and in Chapter 7, 7.7.4, concerning the connection with characteristic functions and Khintchin's theorem.

6.5.8. *Probability distributions and distributional knowledge*. We are now in a position to summarize the conclusions we have reached as a result of following through the weak formulation in a coherent fashion, and also the conventions that have proved necessary in order to make the formalism and the language conform to the requirements of the formulation. In fact, we shall not merely provide a summary, but also fill in some more details, mentioning in an integrated manner certain points hitherto made only incidentaily: in this way, we shall build up the complete picture.

The distinction, originally presented as if it were a small difference in attitude, between a complete distribution, attached to a random quantity and containing all the information about it, and a distribution function as a mathematical entity, useful for providing a partial indication of the form of a random quantity, is now much more sharply drawn. We have seen, in fact, a number of ways in which the latter form is incomplete and not sufficiently informative; this became clear as we proceeded to 're-evaluate' the notion.

Distributional knowledge, as we introduced it (in a way we considered appropriate to make of it an instrument whose range of application was properly defined), is sufficient to obtain a description of the image of a 'distribution of probability mass' within well-determined 'realistic' limits. One can ask how much mass is contained in an interval (but without being able to state precisely whether the mass adherent to the end-points is inside or outside the interval, and with no possibility of saying anything with respect to a set having a complicated form, or not expressible in terms of intervals). One can ask for the mean value of any continuous function with respect to the mass distribution (but not for functions in general, unless one assumes some further conditions). Nothing, however, can be known precisely concerning which points are possible and, without this knowledge, we cannot even say whether or not the mean value of the distribution is the prevision of an X having that particular distribution function.

To summarize: distributional knowledge is only partial, and has to be made precise before it provides complete knowledge. By making it precise, one can obtain many different probability distributions from it; they all have in common, so to speak, those features that are apparent at first sight, without examining the details more closely under a microscope.

Given this analysis, one can now pick out those properties which the strong formulation obtains from the distribution function by virtue of the assumption of countable additivity. These properties might or might not hold (by chance), and might also hold for nonmeasurable sets or functions (should these be of interest). Above all, one needs to state precisely what one means by 'possible points'.

In order to avoid any misunderstandings or ambiguity, and to pay close attention to the distinctions we have drawn, it would be better if we reserved the term '*probability distribution*' for the complete distribution, $F_{\mathscr{G}}$, and always used '*distribution function*' for what, in an abstract sense, should be called 'the equivalence class of all the probability distributions which are the same if we confine ourselves to $F_{\mathscr{R}}$' (to put it briefly, and more intuitively, 'when we look at them with the naked eye'), and which, in the final analysis, can be said to be $F(x)$. This would be (perhaps?) a little overdone, compared with the standard practice of always saying 'distribution'. At times (when it seems necessary to emphasize the point), we shall be more precise and say 'in the sense of a distribution function'; however, it will generally be left unstated, and clear from the context. What is important is that the reader always bears in mind 'as a matter of principle' that it is necessary to draw a distinction between those things which depend only on $F(x)$, and those which do not.

6.5.9. *A decisive remark*. We have been led, for various reasons, to rule out the assumption of countable additivity. Although it is not directly relevant to our specific purpose, we ought perhaps to give some thought to the reasons why most people are quite happy to accept this assumption as not unreasonable.

Leaving aside the question of analytic 'convenience', seen within the Lebesgue framework (which, in any case, appeared on the scene afterwards), I think the reason lies in our habit of representing everything on the real line (or in finite-dimensional spaces), and in the fact that the line (and these kinds of spaces) does not lend itself to being intuitively divided up into pieces other than those which get included 'by the skin of their teeth'.

To see this, note that the partitions actually made are those which are easiest to make: the 'whole' (length, area, mass etc.) is divided into a finite number of separate parts, with an epsilonth left over; in order to obtain an infinite partition, one carries on

dividing up that epsilonth. If one has to share out a cake among n persons, one could always give $\frac{1}{2}$ to the first one, $\frac{1}{4}$ to the second, $\frac{1}{8}$ to the third, ..., $\left(\frac{1}{2}\right)^{n-1}$ to the last two; if there were a countable infinity of persons, one could cope with them all by this method. But would they be satisfied? Protests would quite likely arise by the time one reached $n = 3$, and, as one proceeded, the number who came to regard this as some kind of practical joke rather than a 'genuine' method of distribution would increase, as would, quite understandably, their anger.

A 'genuine' method, in this sense, for subdividing an interval into a countable partition, is that used by Vitali, in proving the theorem we referred to earlier. The set I_h is formed from points of the form $a + r_h$, where $r_0 = 0$, r_1, r_2, ..., r_n, ... are the rationals (ordered as a sequence), and the a are the irrational numbers of I_0, chosen so that one and only one representative from each set of irrationals which differ among themselves by rationals is taken. This example has a pathological flavour, however, as a reshuffing of the points, not to mention its evident appeal to the axiom of choice.

In contrast, if we considered a space with a countable number of dimensions, the matter would be obvious. If a point is 'chosen at random' on the sphere $\sum_h x_h^2 = 1$ in the space of elements with countably many coordinates x_h, all zero except – at most – a finite number, then there is equal probability (zero – see Chapter 4, and the appendix, Section 18) that any of the half-lines x_h (positive or negative) will be 'the closest half-line'. Leaving aside the 'random choice', the countably many 'pieces' of the sphere, I_h' and I_h'', defined by 'x_h is the greatest coordinate – in absolute value – and is positive (I') or negative (I'')' are entirely 'symmetric' and 'intuitive' (the number of dimensions is, of course, so much greater than three).

The essence of the remark can be put, rather more briefly, in another way. By a set of measure zero, the currently fashionable measure theory means a set that is *too empty* to serve as an element of a countable partition. This is a direct consequence of imposing countable additivity as an axiom. This implies, in fact, that a union of a countable number of sets of measure zero (in the Lebesgue sense) is still of measure zero. It is no wonder that in such a docile set-up any kind of process consisting in taking limits is successful, once all the necessary safety devices have been incorporated in the definitions!

6.6 The Practical Study of Distribution Functions

6.6.1. What we are going to say here holds for any kind of distribution: one can, if one wishes to form a particularly meaningful image, think of mass distributions; or (bearing in mind that we are dealing with the 'distribution function') one can think in terms of the probability distribution, which is the thing we are specifically interested in. It will, however, be most useful, particularly for the more practical aspects, to think mainly in terms of the statistical distribution.

In studying a distribution, we may, roughly speaking, distinguish three kinds of ideas and tools:

descriptive properties,
synthetic characteristics,
analytic characteristics.

6.6.2. Many of the properties already mentioned are *descriptive properties*. As examples, we have the following: whether a distribution is bounded or not; proper or improper; whether $F(\square)$ is finite, infinite (negative or positive) or indeterminate ($\infty - \infty$); whether or not there are masses of each type A, B and C (6.2.3), and, in particular, in case A, whether the density is bounded, continuous or analytic; whether this density (or, in case C, the concentrated masses, for example with integer possible values) is increasing, decreasing, or increases to a maximum and then decreases (*unimodal* distribution), or whether the behaviour is different again (for example, *bimodal* etc.); whether the distribution is symmetric about the origin ($F(-x) + F(x) = 1$) or about some other value $x = \xi$ ($F(\xi - x) + F(\xi + x) = 1$; if the density exists, $f(\xi - x) = f(\xi + x)$, and, in particular, $f(-x) = f(x)$ if $\xi = 0$).

We could continue in this way but it is sufficient to say that one should note how useful it can be to provide sketches showing these various aspects. Sometimes these alone will be enough for one to draw simple conclusions; more frequently, they provide useful background knowledge to be considered along with quantitative data.

6.6.3. In order to be able to interpret what we shall say later by making use of various graphical devices (and, in this way, to better appreciate both the meanings of the different notions, and the properties and particular advantages of each method), we will mention briefly the principal graphical techniques used.

We shall present them using the language of the statistical distribution (for N 'individuals') but they are completely general (if we consider the cases of continuous distributions as covered by taking N very large, or, in mathematical terms, mentally taking the 'limit as $N \to \infty$'). For convenience, we shall only deal with bounded distributions over the positive real line ($F(0) = 0$, $F(K) = 1$, $K = \sup F < \infty$). This will be useful for fixing ideas, necessary for some of the points we shall make, and quite sufficient to show how the same things go through in the general case, with appropriate modifications.

The *graph of the distribution function*, $y = F(x)$, is given in Figure 6.2a; in the statistical case this becomes a step function (which in the limit is a *curve*), called the *cumulative frequency curve*, with a step of $1/N$ at each point x_h, the value, for the hth of the N individuals, taken by the quantity under consideration (for example: age, height, income etc.). $F(x)$ gives the frequency, that is the percentage,[18] $n(x)/N$, of the individuals (out of the total of N) for whom the quantity has a value not exceeding x.

As we already pointed out (6.2.5), the 'individuals' must sometimes be counted with different 'weights' p_h (instead of each with $1/N$); it could also happen that several individuals may have the same value x_h (and we then have a mass at that point of, $\sum_k p_k (x_k = x_h)$, or, in particular, n/N if the masses are equal and n values coincide). We shall concentrate on the simplest case, however, in order to fix ideas concerning certain aspects of importance, without prejudicing the extension to the more general case.

The graph of the inverse function, $x = F^{-1}(\gamma)$, which we considered already in Section 6.2.6 (Figure 6.2b), is not widely used. It is, however, a meaningful concept known as the *gradation curve* (Galton); its interpretation is best illustrated in the case

18 By 'percentage', we mean the proportion (not the proportion multiplied by 100 as is customary): in other words, 27% = 0·27, 27·58% = 0.2758 etc. Nothing is altered (we could mention that this way of *writing* it is convenient in that it avoids zeroes on the left, and is more expressive when it comes to *reading* it): the symbol % is a conventional form of '/100' (divided by 100), as a right operator on any number.

of heights – it is the profile obtained by lining up the individuals in increasing order of height (a kind of 'Right dress!').

When income is the quantity of interest, one could think, for instance, of a pile of equal coins rather than of the individuals. This image is useful for clarifying the concept required in cases like the present one, where an obvious meaning attaches to the *sum* of the x_h values of the various individuals; here, the total income of a certain group of individuals. The area under the curve, and relative to a given interval $\gamma' \leqslant y \leqslant y''$, represents the total income (reduced, on that scale, from 1 to $1/N$) of the individuals belonging to the group of those for whom the percentage point of 'the least rich among them' lies between γ' and y''. In any case, dividing by the length of the segment, one always obtains the mean value (arithmetic mean) of that group of individuals, and this also makes some sense in the case of age and height etc., although the meaning is rather one of convention, since the sum does not have a straightforward interpretation. In any interval (and, in particular, for the whole interval [0, 1]) the mean value is, therefore, the height of the rectangle of equivalent area (in other words, in more visual terms, leaving equal areas above and below).

In those cases where the sum has an obvious meaning (as in the case of income), a third graphical device is also useful and meaningful. It is known as the 'concentration curve', and is the cumulative version of the previous one (with the total area taken to be unity by convention: e.g. total income = 1). Figure 6.3 shows the *concentration curve* $z = G(y)$ (Lorentz), and the *gradation curve* $x = F^{-1}(y)$ displayed together, with total income and average income, respectively, taken as the units of measure. By definition, $G(y)$ represents the fraction of the total income owned by the fraction y of least wealthy individuals. In the case of a uniform distribution (all incomes equal) the curve would be the diagonal of the square $G(y) = y$; in general, the area between the curve and this diagonal – called the area of concentration – when divided by the maximum possible area, $\frac{1}{2}$ (corresponding to all income in the hands of one of the N individuals, N large) is called the *concentration ratio*, and gives an idea of the inequality of distribution (Gini). At each point, the slope of $z = G(y)$ is given by $x = F^{-1}(y)$; the mean corresponds to the point of maximum distance from the diagonal (where $G'(y) = 1$, we have a tangent parallel to the diagonal).

6.6.4. The representation by means of the *density curve* is widely used; in the statistical case this is called the *frequency* curve. It is this representation which best shows up the features of behaviour that we were discussing earlier.

We must point out, however, that the density is often (and, strictly speaking, in the statistical interpretation always) a fiction, or a mathematical idealization. Any actual statistical distribution (with a finite number of individuals, N) must be discrete: we either have N masses p_h (possibly equal – $p_h = 1/N$ – possibly not) with $\sum_h p_h = 1$, or fewer than N if several individual values are equal. Even in the actual case of a distribution of mass, we would find similar discontinuities once we descended to the atomic scale, or even indeterminacy because of thermo-agitation and so on, which would prevent us localizing the masses precisely.

In actual fact, even in physics, the density is acknowledged to be a sensible tool if we consider the ratios of mass/volume for neighbourhoods of a point which are not too large, so that macroscopic inhomogeneity has little effect, and not too small, so that the effects of structural discontinuity are avoided. In any case, if we make the transition from step

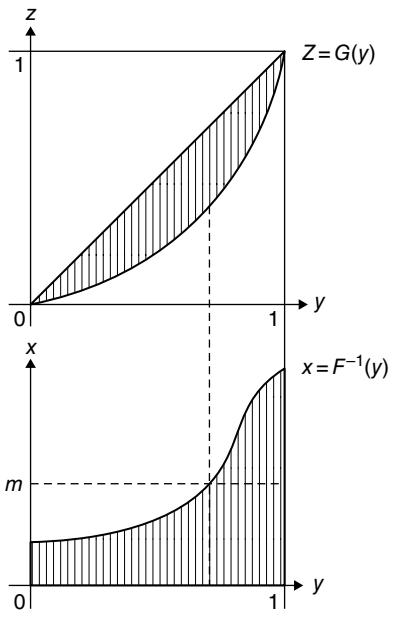

Figure 6.3 The concentration curve $z = G(y)$; for example, in the case of incomes, to the fraction y of the least wealthy, there corresponds the fraction $G(y)$ of total income, which is represented on the graph below (the gradation curve: see Figure 6.2b and the discussion in 6.2.6) by the fraction of the total area to the left of y; that is, including all incomes $\leqslant x = m$ dz/dy (m = average income). Observe, in particular, that $x = m$ at the point where the curve $z = G(y)$ has slope = 1 (the tangent is parallel to the diagonal; it is therefore the point of maximum distance from the diagonal). The diagonal, $z = y$, corresponds to the case of equal distribution; in all other cases, we must have $z < y$, z increasing and concave.

function to distribution function without attributing to the latter any unnecessary irregularities of slope, then $f(x) = F'(x)$ can, to a large extent, be considered as determined. On the other hand, the curve is sometimes *smoothed;* that is, modified in order to simplify it, possibly into a more tractable analytic form, more or less of a standard type.

It is sometimes stated, in this context, that one is attempting to remove '*accidental* irregularities'. This, however, can only be done from a probabilistic angle and in the necessary depth. For this reason, we shall not go into the question here. anything we might say would only tend to give rise to superficial and misleading ideas, which can come about easily enough, even without our saying anything (we shall come back to this in Chapters 11 and 12; we hinted at the underlying idea in Chapter 5, 5.8.7).

The most elementary and, at the same time, the best way of introducing the density in practice (and of constructing the density curve) consists of considering the *average density* over intervals of some appropriate subdivision (neither too coarse nor too fine, for reasons stated already). Unless there is any reason to do otherwise, we usually take equal subintervals (for convenience). The average density in the general interval $[\xi_i, \xi_{i+1}]$, is the incremental ratio of $F(x)$, $[F(\xi_{i+1}) - F(\xi_i)]/(\xi_{i+1} - \xi_i)$. Figure 6.4, formed by rectangles whose bases are the subintervals, and whose height is the average density, is called the *histogram*[19] (sometimes called a column diagram). Here also, by smoothing, one can pass to a continuous *curve*.

6.6.5. The *synthetic characteristics* are the quantitative aspects, which often provide useful information, enabling us to find out all we need to know about the distribution in

19 Note that it is essential to indicate the subdivisions between the rectangles (and that it is not sufficient merely to provide the upper contour). In fact, it is essential to distinguish the case of two (or more) consecutive rectangles of equal height from the case of a single rectangle given by their union. In the first instance there is more information, since we know that the average density is the same in the different subintervals.

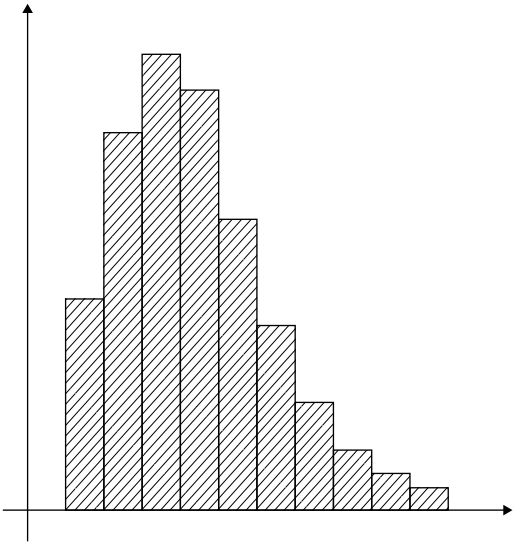

Figure 6.4 An example of a histogram. (It represents the distribution of families in Italy in 1951, according to the number in each.)

so far as it relates to a particular problem. It is sufficient to recall Chisini's definition of a *mean* (Chapter 2, Section 2.9), in order to understand how the knowledge of a 'mean' of a distribution can meet our needs. Often, this will be the mean value (arithmetic mean), given by $F(\square)$, or some other *associative* mean, $\gamma^{-1}F(\gamma)$, with γ increasing, corresponding, in the probabilistic interpretation, to the *prevision*, $\mathbf{P}(X) = F(\square)$, or, more generally, to *the γ-prevision:*

$$\mathbf{P}_\gamma(X) = \gamma^{-1}\left[\mathbf{P}\left(\gamma(X)\right)\right]. \tag{6.3}$$

Sometimes, in addition to the mean (or prevision), one requires the *separation*, $X - \xi$, or the *deviation*, $|X - \xi|$ (the absolute value of the separation), of X from a given pointe ξ (which may be anything). On occasions, it will be particular choices of ξ which are important, as we have already seen in the case of the standard deviation – the quadratic prevision of $|X - \xi|$ with $\xi = \mathbf{P}(X)$ – because it is with this choice of ξ that it assumes its minimum value and maximum significance. Leaving aside the probabilistic interpretation, to consider the separation is simply to consider shifting (from 0 to $-\xi$) the origin of the distribution; to consider the deviation is to turn over that part of the distribution on the negative axis and superimpose it on the positive axis.

Finally, we note that there are other synthetic characteristics which cannot be viewed as means (at least, not without distorting their meanings).

6.6.6. According to the purpose in hand, one can distinguish between measures of *location* and measures of *dispersion* (or spread), which are useful in giving some idea of 'whereabouts' the distribution tends to be concentrated, and 'to what extent' it is concentrated (these are often the two features of greatest interest). Other characteristics which one occasionally attempts to measure by some kind of indices are, for example, the *asymmetry*, the '*kurtosis*' and so on. A brief remark or two will suffice.[20]

20 For a more extensive treatment, see M.G. Kendall, and A. Stuart, *The Advanced Theory of Statistics* (3rd edn), vol. I, Griffin, London (1969), pp. 32–93.

The most meaningful measures of *location* are, generally speaking, the means (in which the Chisini sense; precisely because of the property expressed by his definition). Most often, however, one is interested in measures which behave sensibly under *translation* (and we implicitly mean *homogeneous:* in other words, if X transforms to $aX + b$, the measure is multiplied by a and increased by b). In general, this property does not hold: for example, among associative means only the arithmetic mean has the property.[21]

Examples of measures of location which do have the required property are the commonly used *median* (or median value) and *mode* (or modal value) of a distribution.

The *mode* is the value for which the density is a maximum. It is clearly defined and meaningful in the case of distributions whose densities have regular behaviour, and which are unimodal (that is, have a unique maximum), especially when defined in terms of simple functions. The more we depart from such well-behaved situations, the less clearly defined and meaningful it becomes.

The *median* is the central value of the distribution, the value which splits it in half; that is, such that $F(x) = \frac{1}{2}$ (or, more explicitly, $x = F^{-1}(\frac{1}{2})$). It is of the value which has the property of minimizing $\mathbf{P}[|X - \xi|]$, the prevision of the deviation.[22]

The median is a special case – the most important- of a *positional value*, or *quantile*, of a distribution. The definition of the *p-quantile* ($0 \leqslant p \leqslant 1$) follows along the same lines; $x_p = F^{-1}(p)$, that is, the value which divides up the distribution into a mass p on the left, and $1 - p$ on the right. For $p = 0$ and $p = 1$, we have inf X and sup X (making the natural convention of choosing one of these values rather than any value <inf X or >sup X). These values have the translation property, but are not suitable (for $p \neq \frac{1}{2}$) as really meaningful measures of location; they are useful as 'milestones', well suited to describing the distribution in terms of intuitive subdivisions, especially when considering *quartiles* ($p = \frac{1}{4}$ or $p = \frac{3}{4}$), *deciles* and *centiles* (p multiples of $\frac{1}{10}$ or $\frac{1}{100}$), or for furnishing measures of dispersion (as we shall see).

In the case of measures of *dispersion* (or, if looked at in the opposite sense, measures of *concentration*), it will also prove important to consider a *homogeneity* property (similar to the translation property considered above). For the most important measures, when we consider $aX + b$ the measure is multiplied by a (and b has no effect).

Let us consider the special case of a distribution transformed into its '*normalized*' (or standardized) form, by taking the mean value as the origin, and the standard deviation as the unit ($m = 0$, $\sigma = 1$). If we denote by α^* the index for the normalized distribution, then, after transformation, the translation property would lead to $\alpha = m + \sigma\alpha^*$, and the homogeneity property to $\alpha = \sigma\alpha^*$. If $\alpha = \alpha^*$ (in other words, invariance under translation and change of scale) the index could be called *morphological*, because it expresses a characteristic of the form of the distribution, that is, of the *kind* of distribution (this terminology is often useful for denoting all those distributions which differ from each other only by changes of origin and scale; in other words, the $F(ax + b)$ for given F and

21 It holds for the others if the scale is transformed by $y = \gamma(x)$.

22 This is obvious if one thinks about it. Shifting ξ to $\xi + d\xi (d\xi > 0)$ increases by $d\xi$ the deviation for all masses to the left of ξ, and decreases by the same amount the deviation for those on the right. It is therefore sensible to move towards the median, at which point the masses on the left and right are equal. This property (with an appropriate modification) allows us to eliminate the indeterminacy which occurs in $F(\xi) = \frac{1}{2}$ throughout some interval. One can define (D. Jackson, 1921) the median as the limit as $\varepsilon \to 0$ of $\xi(\varepsilon)$ = the value at which the prevision of the deviation to the power $1 + \varepsilon$ ($\varepsilon > 0$) is minimal.

any a and b; sometimes, we are limited to $a > 0$ and/or $b = 0$). Observe that we carried out the normalization using m and σ, but this is by no means the only possibility, nor is it even always possible (σ may be infinite, or m indeterminate); we used this method because it is the most common, and the most useful from several points of view. As an example of the other possibilities, we mention the possibility of taking the *median* and the *interquartile range*, in place of m and σ (this has the advantage that it is always meaningful, and avoids the oversensitivity of σ to the 'tails' of the distribution; its disadvantage is that it is rather crude).

Examples of morphological properties are provided by *asymmetry* and *kurtosis*, for which one can take as indices the cubic and quartic means of the separation $-\mathbf{P}[(X-m)^n]^{1/n}$ for $n = 3$ and $n = 4$, respectively, divided by σ.[23] The first index is equal to 0 in the case of symmetry (or of deviations from symmetry which cancel each other out),[24] and is positive or negative according to whether the left-hand or right-hand tail is more pronounced. Kurtosis, measured by the second index, is the property of whether the density is sharp or flat around its maximum, and its main use is in discovering whether a density which appears to be *normal* (see 6.11.3) is, instead, *leptokurtic* or *platykurtic;* that is, more peaked or more flat than it should be around the maximum. The index given distinguishes between the three cases depending on whether it is $=, >, <^4\sqrt{3}$.

Let us now go back to the case of dispersion and mention, in addition to the mean deviations (from m or any other value), the means of the differences, $\mathbf{P}[|X - Y|]$ or $\mathbf{P}_\gamma[|X - Y|]$, where X and Y are independent random quantities having the distribution under consideration. The *mean difference*,[25] $\mathbf{P}[|X - Y|]$, is expressible (for distributions on the positive axis) in terms of the area of concentration (see 6.6.3, Figure 6.3); the *quadratic mean difference*, $\mathbf{P}_Q[|X - Y|]$, does not give us anything new, it is clearly equal to $\sqrt{2}\sigma$ ($\sqrt{(\sigma^2 + \sigma^2)}$). Other indices can be set up in terms of quantiles: the *interquartile range* and the *intersecile range* are, respectively, the differences between the quantiles with $p = \frac{1}{4}$ and $p = \frac{1}{3}$, and with $p = \frac{1}{10}$ and $p = \frac{9}{10}$; the limits, $p = 0$ and $p = 1$, give the range of the distribution; sup – inf.

A somewhat different concept of dispersion lies behind the function $l(p)$, $(0 \leqslant p \leqslant 1)$ defined by $l(p)$ = 'the minimum length of a segment containing mass (probability) p' = $\inf\{\lambda \sup_x[F(x + \lambda) - F(x)] \geqslant p\}$. Clearly, $l(p) = 0$ for $p \leqslant$ 'the maximum jump' (the maximum probability concentrated at a point; in particular, if there are no concentrated masses then $l(p) = 0$ only when p = 0); $l(p)$ is increasing, and tends to the range of the distribution as $p \to 1$. If $l'(0) = c > 0$, the distribution has a bounded density, and its maximum is $1/c$ (and conversely).

---

23 More usually, powers are used: it seems preferable and more meaningful to take ratios of means of dimensionality 1 with respect to the variable.

24 Observe how this cancelling out depends on the particular choice of the index. In general, any index which translates an essentially qualitative property into a quantitative measure introduces a degree of arbitrariness. One should take account of this both by exercising caution in interpreting the conclusions, and also by avoiding abstract verbal discussions concerning the 'preferability' of various indices; this question should, if at all, be examined in relationship to the concrete needs of the problem.

25 In the case of the statistical distribution (with N individuals) one considers mean differences *with* and *without repetition*. The latter implies that one excludes X and Y referring to the same individual (excluding the fact that it can be drawn twice) and the index is then multiplied by $N/(N - 1)$. In fact, the probability of a repeat drawing is $1/N$; hence, we have '*index with*' = $(1 - 1/N)$. "*index without*" $(1/N).0$ (0 being the difference between X and Y when they coincide).

6.7 Limits of Distributions

6.7.1. We have had occasion to note that certain properties and synthetic characteristics of the distribution function are rather insensitive to 'small changes in the form of the distribution', while others are very sensitive. To make this more precise, we must first say what we mean by a 'small change'; at the very least, this implies saying what we mean by a sequence of distributions, $F_n(x)$, tending to a given distribution $F(x)$ as $n \to \infty$. Better still, when this is possible, it means defining a notion of 'distance' between two distributions, allowing us to recast $F_n \to F$ in the form dist $(F_n, F) \to 0$.

Fortunately, there is little doubt about what form of convergence is appropriate in the case of proper distributions (and we shall limit ourselves to this case). To say that $F_n \to F$ will always mean convergence of $F_n(x)$ to $F(x)$ at all continuity points of F (or, alternatively, convergence of $F_n(\gamma)$ to $F(\gamma)$ for every bounded and continuous γ). An equivalent formulation is expressed by the condition:

given any $\varepsilon > 0$, the inequalities

$$F(x-\varepsilon)-\varepsilon \leqslant F_n(x) \leqslant F(x+\varepsilon) + \varepsilon \quad \left(-\infty \leqslant x \leqslant \infty\right) \tag{6.4}$$

are satisfied for all n greater than some N.

A condition of this form makes it evident that the smallest value of e for which it holds can be defined as the *distance*, dist (F_n, F), between F_n and F (geometrically, this is the greatest distance between the curves $y = F_n(x)$ and $y = F(x)$ in the direction of the bisector $y = -x$). We shall not prove this; we merely observe that this corresponds to the idea that a given imprecision is tolerated not only in the ordinates (a small change in the mass, in the probability), but also in the abscissae (small changes in the position of the mass, even the concentrated mass).

It often happens that a sequence F_n does not tend to a particular distribution F, but only to a distribution of *the same kind* as F (as defined in 6.6.6). In other words, $F_n(a_n x + b_n)$ tends to F if we *choose* the constants a_n and b_n in an *appropriate manner*. The most common case is that of the normalized distribution $F_n([x - m_n]/\sigma_n)$ (with $a_n = 1/\sigma_n$ and $b_n = -m_n/\sigma_n$), but this is not the only one, and is not always applicable, even when all the variances (of the F_n and of F) are finite and convergence to F occurs (by choosing the constants differently).[26]

6.7.2. We can straightaway make some important points.

Every distribution can be approximated to any desired degree by means of discrete distributions, or by means of absolutely continuous distributions.

It suffices to observe that this follows, for example, if we set

$$F_n(x) = \text{the largest multiple of } 1/n \text{ which is less than } F(x) + 1/2n, \tag{6.5}$$

---

26 The masses which move away (as n increases) and which die away (as $n \to \infty$) without changing the limit of the distributions may, for example, change the σ_n.

Example. Let F_n have masses $\frac{1}{2}(1 - 1/n)$ at ± 1 and masses $\frac{1}{2}n$ at $\pm n$; we have $\sigma_n \sim \sqrt{n} \to \infty$; the normalized F_n would have two masses $\sim \frac{1}{2}$ at $\pm x_n$, $x_n \sim 1/\sqrt{n} \to 0$ (and two which become negligible) and would tend to a distribution concentrated at 0; the F_n (unnormalized) tend, on the other hand, to F, with masses $\frac{1}{2}$ at ± 1.

or, respectively,

$$F_n(x) = \int_0^1 F\left(x + u/n\right) \mathrm{d}u, \tag{6.6}$$

from which it follows that

$$f_n(x) = F_n'(x) = n\left[F\left(x + 1/n\right) - F(x)\right] \leqslant n. \tag{6.6'}$$

As a result :

> *A property which has been established only for discrete distributions* (*or only in the absolutely continuous case, or simply for cases with bounded density*) *holds for all distributions if that property is continuous* (a property is *continuous* if it holds for F whenever it holds for the F_n such that $F_n \to F$).

It is easy to show that continuity usually holds for most of the properties that are required. It is much less long-winded to write out the proof (even if it follows the same lines) in one or other of the special cases, whichever is convenient for our purpose.

It is useful to bear in mind that in order for a sequence F_n to *be convergent* (assuming that the F_n tend to a proper limit F) *it is necessary that the F_n be equally proper* (in the sense that $F_n(x) - F_n(-x)$ tends to 1 as $x \to \infty$, uniformly with respect to n); and, conversely, that this condition is sufficient to ensure *the sequence F_n, or at least a subsequence, tends to a proper limit distribution.* (Ascoli's theorem).

6.8 Various Notions of Convergence for Random Quantities

6.8.1. In the most natural interpretation, the notion of convergence deals with *sequences* of random quantities. However, although for the sake of simplicity we shall deal with sequences $X_1, X_2, \ldots, X_n, \ldots$ ($n \to \infty$), nothing would be altered were we to deal with X_t with $t \to t_0$ (real parameter), or, similarly, with X_t associated with elements t of any space whatsoever (in which $t \to t_0$ makes sense). Instead of a sequence, we might be dealing with a series (but this amounts to the same thing when we consider the sequence of partial sums); instead of a random quantity, we might be dealing with random points in general (for example, 'vectors' or n-tuples of random quantities), provided that in these spaces the concepts involved also make sense.

Here we are merely concerned with setting out the basic ideas, and noting, in particular, the numerous points at which the *weak* conception, to which we adhere, leads to formulations and conclusions different from those usually obtained as a result of following the *strong* conception.[27]

---

27 It is not a question, of course, of declaring a preference for weak convergence or strong convergence (although the identity of the terminology does reflect a relationship between the concepts). In both the weak and strong formulations these and other notions of convergence exist, and each might present some difficulties of interpretation in one or the other formulation.

6.8.2. In the first place, it is possible to have definite convergence, uniform or nonuniform, either with a definite limit or not; by *definite* we mean independent of the evaluation of the probabilities; in other words, something that can be decided purely on the basis of what is known to be *possible* or *impossible*.

As an example of definite, uniform convergence to a definite limit, consider the total gain in a sequence of coin tosses (Heads and Tails). A 'success' is defined by the occurrence of a Head, or by 100 consecutive Tails following the last success; the gain is $\left(\frac{1}{2}\right)^n$ for the nth success, and 0 for a failure. The total possible gain is 1, and it is certain that after at most $100n$ tosses the first n terms will have been summed.

Definite, uniform convergence, but to an uncertain (random) limit, occurs in a sequence of coin tosses if the successive gains are $\pm\frac{1}{2}, \pm\left(\frac{1}{2}\right)^2, \pm\left(\frac{1}{2}\right)^3, \ldots, \pm\left(\frac{1}{2}\right)^n, \ldots$ (+ for Heads, – for Tails); the remaining gain after n tosses is (in absolute value) certainly $\leqslant(½)^n$ but the limit could be any number between −1 and +1.

In the following example, convergence is definite, non-niform, and may be either to a definite or to an uncertain (random) limit. We have an urn containing $2N$ balls, a finite number, but for which no upper bound is known. There are $N + X$ white balls and $N - X$ black balls, where $X = x$ may be known (certain; e.g. $x = 0$), or may be unknown (e.g. any number between ±100).

The balls are drawn without replacement, and the gains are ±1(+ for white, – for black). After all drawings, the gain will be $2X$ and will remain so thereafter (we assume, to avoid nuances of language, that when the urn is empty some other fictitious drawings, all of gain 0, are made). The limit is $2X$, either known or unknown, but objectively determined right from the very beginning.

So far, probabilities have not entered onto the scene (nor, therefore, have probabilistic kinds of properties, like stochastic independence). One might ask, however, whether knowing the limit X (as a certain value, x), or attributing to it some probability distribution $F(x)$ (if it is uncertain), imposes some constraints on the evaluations of the probability distributions $F_n(x)$ of the X_n (or conversely: it amounts to the same thing).[28]

In the case of uniform convergence the answer is yes: if we are to have $|X_n - X| < \varepsilon_n$ with certainty, then F_n and F must be 'close to each other' in the sense that $F_n(x - \varepsilon_n) < F(x) < F_n(x + \varepsilon_n)$ (and conversely: $F(x - \varepsilon_n) < F_n(x) < F(x + \varepsilon_n)$). In particular, if $X = x_0$ with certainty, we must have $F_n(x_0 - \varepsilon_n) = 0$, $F_n(x_0 + \varepsilon_n) = 1$. When we are dealing with non-uniform convergence, this does not hold in general (unless we accept countable additivity). In the example of the urn, if $2N$ has an improper distribution (for example, equal probabilities (zero) for each N) then the probabilities of the behaviour of the gain in the first n tosses (however large n is) are the same as for the game of Heads and Tails (whether the difference between the number of white and black balls is known, e.g. = 0, or bounded, e.g. between ± 100 with certainty). Whatever happens, until the urn is emptied (and we know that there is no forewarning that this is about to happen) nothing can be said about the limit (if it is not already known), and knowledge of this limit (if we have it) does not modify the F_n.

6.8.3. Notions of convergence in the probabilistic sense carry a meaning very different from just saying that (with greater or lesser probability) $X_n \to X$ (in the analytic sense of

28 in general, one should consider the joint probability distribution for $X_1, X_2, \ldots, X_n$, for every n; the mention of this fact will suffice here.

being numbers),[29] and from saying that $F_n \to F$ (this can be true for the distributions of X_n and X, without the latter having anything in common).[30]

We give straightaway the three most important types of convergence.

- *Convergence in quadratic mean.* X_n is said to converge to X in quadratic mean, and we write $X_n \overset{\cdot}{\to} X$, if $\mathbf{P}_Q(X_n - X) \to 0$ as $n \to \infty$ (or, equivalently, if $\mathbf{P}(X_n - X)^2 \to 0$). This notion is the simplest, and the most useful in practice; it is related to what we have already said concerning second-order previsions.
- *Weak convergence* (or *convergence in probability*). X_n is said to converge weakly to X, and we write $X_n \overset{<}{\to} X$, if, for any $\varepsilon > 0$,

  $$\mathbf{P}\left(\left|X_n - X\right| > \varepsilon\right) \to 0 \text{ as } n \to \infty.$$

  More explicitly (in order to make a more clear-cut comparison with the case to be considered next) we can state it in the form: for any given $\varepsilon > 0$ and $\theta > 0$, and for all n greater than some appropriately chosen N, all the probabilities $\mathbf{P}(|X_n - X| > \varepsilon)$ are $<\theta$, or (alternatively) all probabilities $\mathbf{P}(|X_n - X| < \varepsilon)$ are $>1 - \theta$.
- *Strong convergence* (or *almost sure convergence*).[31] X_n is said to converge strongly to X, and we write $X_n \overset{>}{\to} X$, if for any $\varepsilon > 0$, $\theta > 0$, and for all n greater than some appropriately chosen N, we not only have all the probabilities $\mathbf{P}(|X_n - X| > \varepsilon)$ that each deviation *separately* is greater than ε being $< \theta$, but we also have the same holding for the probability of even a single one out of an arbitrarily large finite number of deviations from N onwards (n, $n + 1$, $n + 2$,..., $n + k$, ..., $n + K$; $n \geqslant N$, K arbitrary) being $> \varepsilon$. Expressed mathematically,

  $$\mathbf{P}\left[\bigvee_{k=0}^{K}\left|X_{n+k} - X\right| > \varepsilon\right] < \theta\left(\bigvee_{k=0}^{K} = \max \text{ for } k = 0,1,\ldots,K\right),$$

  or

  $$\mathbf{P}\left[\prod_{k=0}^{K}\left(\left|X_{n+k} - X\right| < \varepsilon\right)\right] > 1 - \theta$$

  ($\prod$ = product (arith. = logical) of the events ($|X_{n+k} - X| < \varepsilon$).)

  Put briefly: the probability of any of the deviations being greater than ε must be $<\theta$; in other words, the probability that they are *all* less than ε must be $> 1 - \theta$.

---

29 In connection with the terminological distinction between *stochastic* and *random* (Chapter 1, 1.10.2), we offer here a remark which seems to clarify the various considerations about the X_n (concerning their 'convergence' in various senses), and at the same time to clarify the terminological question. The fact of the numbers X_n, when they are known, tending or not tending to a limit (in some sense or another; convergence pure and simple, Cesàro, Hölder, etc.) can either be *certain* (true or false with certainty), or *uncertain*, given the present state of information: the convergence is then said to be *random*.

Convergence in the probabilistic sense (either the variants we are going to consider, or others) is called *stochastic convergence* because it is not concerned with the values of the X_n, but with circumstances which relate to the evaluation of probabilities (concerning the X_n and possibly an X, which may or may not be their limit in some sense) made by someone in his present state of information. This is something relating not to the facts, but to an opinion about them based on a certain state of information.

30 A warning against confusing these two notions is necessary, not because in themselves they are open to confusion, but because of the dangers of using inappropriate terminology (such as 'random *variable*': see Chapter 1, 1.7.2 and 1.10.2).

31 A form of terminology which is inaccurate in the weak formulation; see the remark to follow and footnote 29.

Remark. In the strong formulation the definition can be more simply stated by talking of '*all* the deviations from N on', rather than of a finite number (K), however large. From a conceptual viewpoint, the question becomes a rather delicate one because an infinite number of events are involved. As usual, this modification is only admissible if countable additivity is assumed.

6.8.4. *The Borel–Cantelli Lemmas.* For a sequence of events E_i, it is required to provide bounds for the probabilities of having at least one success, or no successes, or at least h successes (that is, if Y denotes the number of successes, $Y \geqslant 1$, $Y = 0$, $Y \geqslant h$); all that can be assumed is knowledge of the $p_i = \mathbf{P}(E_i)$. In the *weak* version, this will only make sense if we limit ourselves to finite subsets (with, of course, the possibility of considering asymptotic results when these subsets cover the whole infinite range). In the *strong* version (as originally considered by Cantelli and Borel, and still standard) the asymptotic results should be interpreted as conclusions about the total number of successes out of the infinite number of events which form the sequence.

For a finite number of events, with probabilities $p_1, p_2, \ldots, p_n$, if we put $\overline{y} = \mathbf{P}(Y) = \sum p_i =$ prevision of the number of successes, we have (unconditionally) an upper bound on the probability of the number of successes:

$$\mathbf{P}\left(Y \geqslant 1\right) = \mathbf{P}\left(\text{event} - \text{sum of the } E_i\right) \leqslant \overline{y}, \mathbf{P}\left(Y \geqslant h\right) \leqslant \overline{y} / h.$$

(In fact, $h\mathbf{P}(Y \geqslant h) = \mathbf{P}[h(Y \geqslant h)]$ and $h(Y \geqslant h)$, which is = 0 if $0 \leqslant Y < h$ and is = h if $Y \geqslant h$, is always $\leqslant Y$: $\vdash h(Y \geqslant h) \leqslant Y$)

We therefore have that if for the sequence E_i the sum of the p_i converges, let us say $\sum p_i = a < \infty$, then $\overline{y} \leqslant a$ for any finite subset, and the previous bounds are valid *a fortiori* (with a in place of $\overline{y}$). One can now say that for any $\varepsilon > 0$, and for $h \geqslant a/\varepsilon$, we have a probability $< \varepsilon$ of obtaining more than h successes among the first K events of the sequence (it does not matter how large K is). In addition, if we only use the bound for $h = 1$, and we start with an n sufficiently large for the rest of the series to be $< \varepsilon (\sum_{i > n} p_i < \varepsilon)$, we can say that the probability of finding even a single success out of K events (K arbitrarily large, but finite) from E_n on is always $< \varepsilon$.

In the strong version we have the following: *if the series of probabilities converges, it is practically certain* (the probability = 1) *that the number of successes is finite.*

This is the Cantelli lemma; the Borel lemma states the converse, but with the additional condition of stochastic independence.[32] In the *strong* version, the divergence of $\sum_i p_i$ implies that the number of successes is infinite; the *weak* version is much the same in this case, because Y, if not infinite, must be a completely improper random quantity (with distribution adherent to $+\infty$).

The bound that is required can be established immediately using the elementary inequality $e^x \geqslant 1 + x$; the probability of no successes in *n independent* events is

$$\begin{aligned}\mathbf{P}\left(Y = 0\right) &= \left(1 - p_1\right)\left(1 - p_2\right)\ldots\left(1 - p_n\right) \leqslant e^{-p_1} \mathrm{e}^{-p2} \ldots \mathrm{e}^{-p_n} \\ &= \mathrm{e}^{-\left(p_1 + p_2 + \ldots + p_n\right)} = \mathrm{e}^{-\overline{y}};\end{aligned}$$

32 It is obvious that this would not hold without any extra condition: think of the case in which the E_i are all incompatible with some E having $P(E) \geqslant a > 0$, such that E implies no successes; i.e. $Y = 0$ (and, in particular $Y_n = 0$ out of the first n of the E_i), so that $P(Y = 0)$ and $P(Y_n = 0)$ are both $\geqslant a > 0$ (instead of = 0 and $\to 0$, respectively). If, however, the series of the $p_i = P(E_t)$ diverges, the E_i cannot then be independent (see the following inequality for $P(Y_n = 0)$).

stated explicitly,

$$\mathbf{P}\big(Y=0\big)\leqslant \mathrm{e}^{-\overline{y}},\ \mathbf{P}\big(Y\geqslant 1\big)\geqslant 1-\mathrm{e}^{-\overline{y}},$$

and, more generally, we have the similar result

$$\mathbf{P}\big(Y\leqslant h\big)\leqslant \mathrm{e}^{-\overline{y}}\left[1+\big(\alpha\overline{y}\big)+\tfrac{1}{2}\big(\alpha\overline{y}\big)^2+\ldots+1/h!\big(\alpha\overline{y}\big)^h\right],\alpha=\mathrm{e}^{\max p_i}.$$

If the series $\Sigma_i\, p_\mathrm{i}$ diverges, $\overline{y}$, relative to the first K events, tends to + ∞ as K increases, and this is also true if we start from the nth event. The conclusion is that there is a probability → 1 of finding at least one success starting from any arbitrary n, and, hence, a number exceeding any bound. Alternatively, this can be established directly from the fact that $\mathbf{P}(Y \leqslant h)$ also tends to 0, for any h.

6.8.5. *A corollary for strong convergence.* In order that strong convergence holds, it is sufficient that the $\mathbf{P}(|Xn - X| > \varepsilon)$ constitute the terms of a convergent series[33] (and do not merely tend to 0, as required for weak convergence). This condition is also necessary if the $|X_n - X|$ are stochastically independent (or if the events $|X_n - X| > \varepsilon$ are). This is seldom so in cases of interest but one can often obtain the negative result by finding a subsequence of terms, which are sufficiently far apart to be 'practically independent', for which the series of probabilities diverges (when we consider something being 'sufficiently independent', we are thinking of some condition or other to be translated into a rigorous form as appropriate for the case in question).

6.8.6. *Relationships between the different types of convergence. Weak* convergence is implied both by *strong* convergence (as is obvious from the definition) and by convergence in *quadratic mean* (by virtue of Tchebychev's inequality, Chapter 4, 4.17.7). Neither of the latter two implies the other.

In addition to convergence in quadratic mean (also known as convergence *in 2nd-order mean*, or in *mean-square*), one also considers, though less frequently, convergence *in pth-order mean* (where p is any positive number), defined by $\mathbf{P}(|X_n - X|^p > \varepsilon) \to 0$; the condition becomes more restrictive as p increases, and always implies weak convergence.

Definite uniform convergence implies all the above.

Convergence of distributions is implied by weak convergence (and so, *a fortiori*, by all the others).

It is sufficient to note that if the random quantities X and Y are 'sufficiently close to each other' in the sense that $\mathbf{P}(|X - Y| > \varepsilon) < \theta$ (for given $\varepsilon, \theta > 0$), then their distributions F and G are 'sufficiently close to one another'[34] in the sense that (for all x) $F(x - \varepsilon) - \theta \leqslant G(x) \leqslant F(x + \varepsilon) + \theta$. In fact, in order that $X \leqslant x - \varepsilon$, it suffices that either $Y \leqslant x$ or $|X - Y| \geqslant \varepsilon$. Expressed mathematically,

---

33 *A fortiori*, it is sufficient that the series $\Sigma\mathbf{P}(X_n - X)$ converges.

34 It is clear that we could define a *distance* between random quantities conforming to this idea (completely analogous to what we did for distributions in 6.7.1): dist (X, Y) = 'the minimum value that can be given to e and ε for which the given condition remains satisfied'. Note that there is a difficulty with regard to the dimensionality (θ is a probability, a pure number, and ε is in general a length): however (as in many such cases, for example the one given in 6.7.1, where this fact was disguised by denoting both θ and ε in the same way, by ε) this difficulty is irrelevant, because changes in 'distance' due to expressing ε in different units, does not alter the thing which interests us; that is, the topology based on 'dist → 0'.

$$\big(X \leqslant x-\varepsilon\big) \leqslant \big(Y \leqslant x\big) \vee \big(|X-Y| \geqslant \varepsilon\big) \leqslant \big(Y \leqslant x\big) + \big(|X-Y| \geqslant \varepsilon\big);$$

taking probabilities, it follows that $F(x - \varepsilon) \leqslant G(x) + \mathbf{P}(|X - Y| \geqslant \varepsilon)$, and the final term is $<\theta$, by assumption. This proves the first half of the inequality; the other half follows by symmetry.

In the case of weak convergence, however we take ε and θ, the inequalities hold for X_n and X from some $n = N$ on, and hence $F_n \to F$.

6.8.7. *Mutual convergence* (or *Cauchy convergence*). Suppose that for a given sequence X_n we know that $X_n - X_m \to 0$ (in some sense) as $m, n \to \infty$: what can be said about the convergence (in the same sense) of X_n to some random quantity X? If we adopt the strong formulation, we can say that such an X exists. For all the types of convergence that we have considered, '*il n'y a pas lieu de distinguer la convergence mutuelle et la convergence vers une limite*' (to quote P. Lévy, *Addition*, p. 58, Th. 18) ['It is not necessary to distinguish between mutual convergence and convergence to a limit'].

The answer is even more conclusively yes if we are dealing with a random quantity which is a measureable function $X(\omega)$ of the points of a space Ω (and, in this case, we should just mention that the various probabilistic notions, and in particular the notions of convergence, reduce to concepts in analysis – apart from changes in terminology: for example, convergence in probability instead of *in measure*; almost certain convergence instead of *almost everywhere*).

Without the assumption of countable additivity, and with no reference to a 'space of points' (see the quotations from von Neumann and Ulam, Chapter 2, 2.4.3), we might well say that an X_n for which, for example, $\mathbf{P}(X_n - X_m)^2 < \varepsilon$ for all but a finite number of X_m, 'represents the limit to within ε'. There is no possibility, however, of thinking of defining X *by the given passage to the limit*.

In order to be able to talk about X, it is necessary that it be a well-defined *quantity*, independently of the incidental fact of whether it is known or not (and then, in this sense, a random quantity). There are various possibilities (which we distinguish for the purpose of giving examples, not because of fundamental differences): X could be random on account of circumstances *logically independent* of the X_n (and therefore, in principle, capable of being measured or known through relevant procedures or information); it could be definable as some function of a finite number of the X_n (as an example, to underline the absence of any restriction on the possibilities, rather than because it makes any sense, one could think of

$$X = \frac{1}{2}\big(X_{1577} + X_{7814}\big) + \pi^{X_{62}}\left(\mathrm{e}^{X_{54}} - \mathrm{e}^{X_{737296}}\right)$$

or anything else that comes to mind), and these also might depend on some further random factors (e.g. on a random quantity Y which may or may not have any connection with the problem); finally, it might depend on all the X_n (and possibly on other things as well; for instance a Y such as we just mentioned).

In particular, it could in this case be

$$X = \begin{cases} \lim X_n \Big(\text{if the sequence of the } \textit{values} \text{ of the } X_n \text{ turn out to be convergent}\Big) \\ 0 \quad \big(\text{otherwise}\big) \end{cases}$$

(and, if one wished, convergence could be taken in the Cesàro sense, or some other). Here too, X is in fact a well-defined quantity (although it can actually only be known after we know the values of all the X_n).

The sentence concerning convergence would only make sense, however, if for such an X, *actually defined independently of the incidental circumstance of what is at present known or unknown*, it were possible to show that, *in the condition of ignorance deriving from these given circumstances*, our present evaluation of probabilities for the X_n and X are such as to imply $X_n \to X$ in some probabilistic sense (quadratic mean, weak, strong, …). On the contrary, we know that this is not the case in general, not even when $\lim X_n = X$, and still less can it be assumed for an undefinable X which has to appear, phantom-like, from the Cauchy property, and then miraculously materialize.

However, mutual convergence (in the weak sense, and *a fortiori* in other, more restrictive, cases) does determine, if not a random quantity X, the limit distribution F. The discussion given above (at the end of 6.8.6) establishes, in fact, that the distributions F_n and F_m, of X_n and X_m, become arbitrarily 'close', and therefore close to one and the same well-defined F, for n and m sufficiently large. In order to be able to state that there exists a limit distribution F such that $F_n \to F$, it is sufficient, for example, to prove that $\mathbf{P}(X_n - X_m)^2 \to 0$ as m and n tend to ∞.

6.8.8. *Zero-one law* (*Kolmogorov*). We must at least give a mention of a phenomenon that was present in the Borel lemma, and is of a general character, constantly cropping up. In order to be brief (since we only want to deal with it in passing), we shall express ourselves in terms of the strong formulation.

Given an infinite number of independent events, E_i, the probability that only a finite number of them occur ($Y < \infty$) is always 1 if the sum of the probabilities converges, and is always 0 if the sum diverges; intermediate probabilities are not possible.

We shall not give a proof, but the main idea is contained in the following: suppose that an event A (such as $Y < \infty$ in the above) is independent of any property A_n which depends only on the first n trials (for example, whether Y is finite cannot be altered by considering a finite number of trials), but is defined, in the limit as $n \to \infty$, by the A_n. Because of independence, $\mathbf{P}(A_n A) = \mathbf{P}(A_n)\mathbf{P}(A)$; taking the limit $A_n \to A$, we have

$$\mathbf{P}(AA) = \mathbf{P}(A) = \mathbf{P}(A)\mathbf{P}(A) = \left[\mathbf{P}(A)\right]^2$$

which implies $\mathbf{P}(A) = [\mathbf{P}(A)]^2$, and hence the only possible values are 0 and 1.

6.9 Distributions in Two (or More) Dimensions

6.9.1. Everything we have said in the one-dimensional case extends straightforwardly to two dimensions (or more: in general, we shall present the extension for $n = 2$, and indicate how to proceed to $n = 3$ etc.). The extension has to be considered now because, even if we only wished to deal with random quantities, as soon as we consider two of them we have to deal with the distribution of the pair (X, Y) as a random point in the plane (x, y). This will not, however, be the only kind of application.

A distribution (always to be interpreted as distribution function) over the (x, y)-plane will always be defined by a *joint distribution function*.

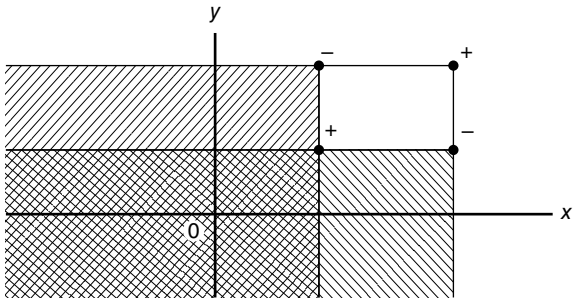

Figure 6.5 Quadrants of the (x, y)-plane, in terms of which the joint distribution function $F(x, y)$ is defined (SW quadrants), and a method of indicating the rectangles with their linear combinations (and, hence, their probabilities in terms of linear combinations of the values $F(x, y)$ at the vertices).

$F(x, y)$ = 'the mass contained in the quadrant SW of the point (x, y)';[35] the mass in the rectangle $x' \leqslant x \leqslant x''$, $y' \leqslant y \leqslant y''$ is then given by

$$F\left(x'', y''\right) - F\left(x'', y'\right) - F\left(x', y''\right) + F\left(x', y'\right); \tag{6.7}$$

see Figure 6.5: rectangle = whole quadrant – hatched quadrants + double-hatched quadrant (since this was taken away twice). The relation can be interpreted as an operation involving masses, or probabilities, or, more basically, a linear combination of the four events 'belonging to the various quadrants under consideration'.

It may be that the masses are concentrated at points, or distributed in an absolutely continuous manner; there are, however, a great variety of intermediate cases (think, for example, of a mass distributed continuously along a line!).

The density (if and when it exists) is given by

$$f\left(x, y\right) = \frac{\partial^2 F}{\partial x \partial y} \tag{6.8}$$

(the limit of the probability given above, with $x'' = x' + h$ and $y'' = y' + k$, divided by the area hk as h and $k \to 0$).

We can define $F(\gamma)$ for functions $\gamma(x, y)$ of two variables, always in the Riemann–Stieltjes sense (and, if γ is not integrable, we have $F^-(\gamma) < F^+(\gamma)$; the probabilistic interpretation is as the bound for $\mathbf{P}[\gamma(X, Y)]$, and, in particular, if $F(\gamma)$ exists, as its evaluation: throughout, the boundedness conditions for the possible values are to be understood, or, if not, the choice of $\breve{\mathbf{P}}$ is understood etc.).

In particular, if $\gamma(x, y)$ represents a set I ($\gamma = 1$ on I and $\gamma = 0$ outside), $F(\gamma) = \mathbf{P}(I)$.

Important examples. If $Z = X + Y$, the distribution function of Z is given by

$$\begin{aligned} \mathbf{P}\left(Z \leqslant z\right) &= F\left(x + y \leqslant z\right) \\ &= F\left(\text{the half} - \text{plane to the SW of the line } x + y = z\right) \end{aligned} \tag{6.9}$$

35 Adopting the practical terminology favoured by economists, we label the 1st, 2nd, 3rd and 4th quadrants as NE, NW, SW and SE (and use these also in referring to directions etc.; the intuitive reference is to a map with N oriented upwards, as usual). Here we implicitly consider F as undefined where it is discontinuous, and so on. Let us simply remark that all the same conceptual details, which we have discussed at length in the one-dimensional case, can be filled in: we shall only do so when some new feature arises, which is something other than a more or less obvious extension of what has gone before.

(in other words, 'the mass contained there'). If Z = *X* Y, we have

$$\begin{aligned}\mathbf{P}\big(Z\leqslant z\big)&=F\big(xy\leqslant z\big)\\&=F\,(\text{the region bounded}^{36}\text{ by the hyperbola }xy=z)\end{aligned}\tag{6.10}$$

(in other words, 'the mass contained there'). If $Z = Y/X$, we have:

$$\begin{aligned}\mathbf{P}\big(Z\leqslant z\big)&=F\big(y/x\leqslant z\big)=F\Big[\big(y\leqslant zx\big)\big(x>0\big)+\big(y\geqslant zx\big)\big(x<0\big)\Big]\\&=F\,(\text{the NW and SE corner regions between the }y\text{-axis and the line }y=zx)\end{aligned}\tag{6.11}$$

(in other words, 'the mass contained there'). If $Z = \sqrt{(X^2 + Y^2)}$, we have

$$\begin{aligned}\mathbf{P}\big(Z\leqslant z\big)&=F\big(x^2+y^2\leqslant z^2\big)\\&=F\big(\text{the disc centred at 0 with radius }z\big)\end{aligned}\tag{6.12}$$

(in other words, 'the mass contained there').

And it would be easy to continue in this manner.

6.9.2. Let us now see how to obtain these results more explicitly. The standard method – integration, using Cartesian coordinates – requires us to make the inequality explicit in terms of one of the variables, *y*, say. In the examples given we have:

$$\begin{aligned}&\textit{sum},\qquad y\leqslant z-x;\\&\textit{product},\ \big(y\leqslant z/x\big)\big(x>0\big)+\big(y\geqslant z/x\big)\big(x<0\big);\\&\textit{quotient},\big(y\leqslant zx\big)\big(x>0\big)+\big(y\geqslant zx\big)\big(x<0\big);\\&\textit{distance},\ |y|\leqslant\sqrt{\big(z^2-x^2\big)}.\end{aligned}$$

In these four cases, the integrals (always either $\int dF$ or $\int f(x, y)\,dx\,dy$) will be

$$\int_{-\infty}^{\infty}dx\int_{-\infty}^{z-x}dy\ldots;\tag{6.9′}$$

$$\int_{-\infty}^{0}dx\int_{z/x}^{+\infty}dy\ldots+\int_{0}^{+\infty}dx\int_{-\infty}^{z/x}dy\ldots;\tag{6.10′}$$

$$\int_{-\infty}^{0}dx\int_{zx}^{\infty}dy\ldots+\int_{0}^{\infty}dx\int_{-\infty}^{zx}dy\ldots;\tag{6.11′}$$

$$\int_{-z}^{z}dx\int_{-\sqrt{(z^2-x^2)}}^{+\sqrt{(z^2-x^2)}}dy\ldots.\tag{6.12′}$$

In general, if $Z = \gamma(X, Y)$ we have $\mathbf{P}(Z \leqslant z) = F(\gamma(x, y) \leqslant z) = F_\gamma(z)$ (say), and if the inequality can easily be made explicit with respect to *y*, obtaining, in the simplest case, $y \leqslant g(x, z)$ (or, possibly, $g_1(x, z) \leqslant y \leqslant g_2(x, z)$), we

36 'Interior' or 'exterior' region, according to whether $z > 0$ or < 0.

shall have

$$F_\gamma(z) = \int_{-\infty}^{\infty} \mathrm{d}x \int_{g_1(x,z)}^{g_2(x,z)} \mathrm{d}y \ldots.$$

Clearly, it may sometimes be more convenient to adopt other coordinate systems (e.g. polar coordinates), remembering, of course, to multiply by the Jacobian.

Let us indicate also how one obtains directly the *density* $f_\gamma(z) = \mathrm{d}F_\gamma(z)/\mathrm{d}z$ (in those cases where everthing goes through smoothly). From the expression for $F_\gamma(z)$, assuming that $F(x, y)$ has a density $f\ (x, y)$, we obtain

$$F_\gamma(z) = \frac{\mathrm{d}}{\mathrm{d}z} \int_{-\infty}^{\infty} \mathrm{d}x \int_{g_1(x,z)}^{g_2(x,z)} f(x,y)\,\mathrm{d}y$$

$$= \int_{-\infty}^{\infty} \mathrm{d}x \left[f\big(x, g_2(x,z)\big) \frac{\partial}{\partial z} g_2(x,z) - \text{the same thing for } g_1 \right].$$

For the examples we have considered, this gives

$$sum: \qquad g_1 = -\infty,\ g_2 = z - x;\ f_s(x) = \int_{-\infty}^{\infty} f(x, z-x)\,\mathrm{d}x; \tag{6.9″}$$

$$\begin{aligned} product: \quad & x < 0 : g_1 = z/x,\ g_1' = 1/x,\ g_2 = +\infty; \\ & x > 0 : g_1 = -\infty,\ g_2 = z/x,\ g_2' = 1/x; \\ & f_p(z) = \int_{-\infty}^{\infty} \frac{1}{|x|}(x, z/x)\,\mathrm{d}x; \end{aligned} \tag{6.10″}$$

$$quotient: \quad \big(\text{as above, with } x \text{ in place of } 1/x\big)$$

$$f_q(z) = \int_{-\infty}^{\infty} |x| f(x, zx)\,\mathrm{d}x; \tag{6.11″}$$

$$\begin{aligned} distance: \quad & -g_1 = g_2 = \sqrt{\big(z^2 - x^2\big)}; \\ & -g_1' = g_2' = z\sqrt{\big(z^2 - x^2\big)}; \end{aligned} \tag{6.12″}$$

$$f_d(z) = \int_{-z}^{z} \frac{z}{\sqrt{\big(z^2 - x^2\big)}} \left\{ f\left(x, \sqrt{\big(z^2 - x^2\big)}\right) + f\left(x, -\sqrt{\big(z^2 - x^2\big)}\right) \right\} \mathrm{d}x.$$

The first example, the simplest, should be noted well, since the case of the sum is basic for most theoretical developments and applications.

We add one last example, where the answer comes out directly: for the *maximum*, $Z = X \vee Y$, the distribution function is given by

$$F(z) = F(z, z) \big(\text{in fact, } (Z \leqslant z) = \big[(X \vee Y) \leqslant z\big] = (X \leqslant z)(Y \leqslant z)\big); \tag{6.13}$$

similarly, for the *minimum*, $Z = X \wedge Y$, the distribution function is given by

$$F(z) = F(z, +\infty) + F(+\infty, z) - F(z, z). \tag{6.14}$$

By means of $F(\gamma)$, we can also, in this case, express various 'synthetic characteristics' of distributions of two variables. For example, for the moments we take $\gamma(x, y) = x^r y^r$ and obtain $M_{r,s} = P(X^r Y^r) = \int x^r y^s \, dF = \int x^r y^s f(x, y) \, dx \, dy$. We have already seen the first- and second-order moments with respect to the origin: $\mathbf{P}(X)$ and $\mathbf{P}(Y)$, the coordinates of the barycentres; $\mathbf{P}(X^2)$, $\mathbf{P}(Y^2)$ and $\mathbf{P}(XY)$, the second-order terms (the moments with respect to the barycentres are

$$\mathbf{P}(X^2) - [\mathbf{P}(X)]^2, \quad \mathbf{P}(Y^2) - [\mathbf{P}(Y)]^2 \quad \text{and} \quad \mathbf{P}(XY) - \mathbf{P}(X)\mathbf{P}(Y),$$

the variances and the covariance). We already know that, in terms of second-order properties, these moments completely characterize the distribution: in particular, we have seen that the cancelling out of the mixed barycentric moment ($\mathbf{P}(XY) - \mathbf{P}(X)\mathbf{P}(Y) = 0$, that is $\mathbf{P}(XY) = \mathbf{P}(X)\mathbf{P}(Y)$, the property referred to as noncorrelation) is a necessary condition for X and Y to be stochastically independent.

6.9.3. *Stochastic independence of random quantities.* The time has come for us to consider the notion of stochastic independence in the context of random quantities (and, essentially, in the most general case, since the delicate issues have a unique character). Up until now, the concept has only been defined (in Chapter 4) for events (4.9.2) and for random quantities with only a finite number of possible values (4.10.1). The extension to the general case is essentially intuitive; we mentioned this (in 4.16.2), where we also pointed out that a detailed and critical approach was required.

The meaning of stochastic independence was: 'that whatever one learns or assumes about X does not modify one's opinion about Y'; put more 'technically', 'every event concerning Y is stochastically independent of every event concerning X'.

Naturally, when it comes to considering n random quantities, these (like events) will *not* be called independent if the independence is merely *pairwise*, but only if each of them is independent of anything one knows or assumes *concerning all the others* simultaneously (that is, of each event concerning all these other random quantities).

Once again we are faced with the question: *which events* do we include in this definition? We might be tempted to say '*all of them*' (and so refer ourselves to $F_{\mathscr{C}}$; but we know that this is a rather unimaginable abstraction); we might say (along with the supporters of the 'strong' formulation) 'all those of the Lebesgue field, or at least the Borel field' (thus referring ourselves to $F_{\mathscr{B}}$; but this runs counter to the objections we have made against countable additivity and the strong formulation); we might limit ourselves to the intervals (and things expressible in terms of them; this leaves us in the field $F_{\mathscr{R}}$). Note, however, that the question does not require a discussion and a decision as to which answer provides the *correct* definition: the best solution would probably be to consider all three definitions (or perhaps none of these), drawing a distinction between 'complete', 'strong' and 'weak' independence. We shall limit ourselves, however, to the weak definition since it is the only one which does not make too unrealistic assumptions about our knowledge. In fact, it is the usual definition, apart from the fact that this

notion has a completed appearance when the unique extension to the Lebesgue field is assumed, along with non-existence outside it.

The assumption that events of the form $X \leqslant x$ are independent of those of the form $Y \leqslant y$ (for any x and y) is sufficient to imply that $F(x, y) = F_1(x)F_2(y)$, where $F_1(x) = F(x, +\infty)$ and $F_2(y) = F(+\infty, y)$ are the distribution functions of X and Y (with the usual qualification of indeterminacy at jump points). It follows immediately that there is also independence for the intervals:

$$\begin{aligned}&\mathbf{P}[(x' \leqslant X \leqslant x'')(y' \leqslant Y \leqslant y'')]\\ &\quad = F_1(x'')F_2(y'') - F_1(x'')F_2(y') - F_1(x')F_2(y'') + F_1(x')F_2(y')\\ &\quad = \left[F_1(x'') - F_1(x')\right].\left[F_2(y'')\right] - F_2(y').\end{aligned}$$

This implies independence for step functions of the single variables x or y, and hence for continuous functions. We conclude that the condition defined by

$$F(x, y) = \text{products of functions involving } x \text{ only and } y \text{ only}, \tag{6.15}$$

is also equivalent to the following condition:

for any product of continuous functions, $\gamma(x, y) = \gamma_1(x)\gamma_2(y)$, we have

$$F(\gamma) = F(\gamma_1)F(\gamma_2), \tag{6.15′}$$

in other words,

$$\mathbf{P}\{\gamma_1(X)\gamma_2(Y)\} = \mathbf{P}\{\gamma_1(X)\gamma_2(Y)\}. \tag{6.15″}$$

6.9.4. Observe, however, how far removed this condition is from the intuitive notion of stochastic independence. We can always assume that the possible points are those of the set of $A_{r,s}$, with coordinates $x_{r,s} = r + s\sqrt{2}$, $y_{r,s} = r + s\sqrt{3}$ (a countable set, since the points are defined in terms of two rationals r and s).

This set is, in fact, everywhere dense in the plane and can be the logical support of any distribution function; in particular, of a distribution which makes X and Y stochastically independent. But, on the other hand, to each possible value for X there corresponds a unique possible value for Y, and conversely (because, given x, there exists at most one pair of rational values r and s giving $x = r + s\sqrt{2}$; if there were another pair, so that $x = r' + s'\sqrt{2}$, we would have $\sqrt{2} = (r - r')/(s - s')$, an absurdity).[37] We can thus have logical dependence (even one-to-one and onto) at the same time as (distributional) stochastic independence. We must bear in mind just how unsatisfactory this definition is from a logical viewpoint, even if it seems difficult to improve on it within the ambit of realistic possibilities.

Remark. Observe that such 'paradoxes' can *also occur in the discrete case*, if the probabilities are thought of not as being concentrated at the points (x_h, y_k), but as *adherent* to them (which is excluded, as in Chapter 4, 4.10.1, if we talk of a 'finite number of

37 From this it also follows that $y = f(x)$ is additive, where it is defined: $f(x' + x'') = f(x') + f(x'')$ but not linear (for $s = 0$, $f(x) = x$; for $r = 0$, $f(x) = \sqrt{(3/2)}x$); and the graph of such a function is dense in the whole plane (see for example, B. de Finetti, *Mathematica logico-intuitiva*, No. 40, 'Sulla propriété distributiva', in particular, Figure 30, pp. 91–92 in the 3rd edn, Cremonese, Rome (1959)).

possible values' – but this subtlety might be overlooked). Therefore, the decision (here in 6.9.3) 'not to give a precise value to F at jump points' is essential.

If a point (x_h, y_k) is not a possible point, but instead (or also) a limit point of a sequence of possible points, each having zero probability, but with positive total probability, a great number of different cases of distributional independence ($p_{hk} = p'_h p''_k$) are possible but other kinds are not (not even logical independence).

6.9.5. On the other hand, we ought to point out that paradoxes (of *nonconglomerability*; see Chapter 4, 4.19.2) arise in connection with 'stochastic independence' without any need to look at pathological examples (or, as some would say, to make them up). The following is a well-known example: if we choose a point 'at random' on the surface of a sphere, equal areas have equal probabilities; and if we happen to know which great circle the point has landed on, then equal arcs will have equal probabilities; if, in addition, we have a system of geographical coordinates (latitude and longitude, say, as on the earth) these coordinates are independent.

In fact, distributional independence holds; the surface element whose latitude lies between ϕ and $\phi + \mathrm{d}\phi$, and whose longitude lies between λ and $\lambda + \mathrm{d}\lambda$, has area $\cos\phi\,\mathrm{d}\phi\,\mathrm{d}\lambda$, and (apart from the normalization constant) this is also its probability. Longitude has a uniform distribution ($1/(2\pi)$ between $\pm\pi$), and latitude has a distribution whose density is given by $f(\phi) = \frac{1}{2}\cos\phi$ (between $\pm\pi/2$); the density for the area (in the λ, ϕ-plane) is the product

$$\frac{1}{2\pi}\cdot\frac{1}{2}\cos\phi = \frac{1}{4\pi}\cos\phi.$$

But then, because of the other assumptions, even if we know the longitude precisely – in other words, the meridian to which the point belongs – the probability distribution of the latitude should always have density $\frac{1}{2}\cos\phi$; on the other hand, since we are dealing with (half) a great circle, the density should be uniform ($=1/\pi$).

The paradox is easily resolved if we argue in terms of 'imprecision'. If, instead of thinking of the point lying exactly on that curve, one thinks in terms of the fact having been ascertained to within some margin of error, however small, one sees that the two answers are coherent. We give two different versions: if the imprecision concerns λ, then, instead of a meridian curve, we have a zone which narrows from the equator to the poles as $\cos\phi$; if, instead, we think in terms of having measured the distance from a plane passing through the centre of the earth (that is, the distance from the great circle) then (finding the distance to be 0) we have a zone of constant width.

It is easy to avoid paradoxes by avoiding any reference to limit-cases, except when considering these explicitly as such (never speak of 'the probability of something conditional on $X = x_0$', but 'conditional on $X = x_0 + \varepsilon$', perhaps giving the limit as $\varepsilon \to 0$). Many authors (the first of them being, I think, Kolmogorov in 1933) explicitly state that the problem only makes sense under this restriction (since, otherwise, conditional probability would formally be given by expressions of the form 0/0). From a theoretical point of view, viewed from the standpoint to which we adhere, such a conclusion seems rather drastic (although it avoids some difficulties, others take their place). Theoretically, it does not seem possible to avoid the necessary comparisons among the zero probabilities which would yield an actual probability for the 'precise' fact, rather than the zero probability usually attributed (see Chapter 4, 4.18); practically speaking, it is convenient to

attempt to use the Kolmogorov limit argument, by considering it in conjunction with what is empirically known about the imprecision (when actually present) and not merely as a convention or a dogma. We shall mention this again later (Chapter 12, 12.4.3).

6.9.6. *Operations on stochastically independent random quantities: Convolutions*. Let us now return to consideration of a random quantity Z = $\gamma(X, Y)$, a function of two other random quantities (as in 6.9.2), in the rather special case where X and Y are stochastically independent. This implies that $F(x, y) = F_1(x)F_2(y)$ and $f(x, y) = f_1(x)f_2(y)$ (if these exist), and we have $dF(x, y) = dF_1(x)\, dF_2(y) = f_1(x)\, f_2(y)\, dx\, dy$.

The fundamental case, which we shall encounter and make use of over and over again, is that of the *sum*, Z = $X + Y$, for which $F(z)$ and $f(z) = F'(z)$ are given by

$$\begin{aligned} F(z) &= \int_{-\infty}^{+\infty} dF_1(x) \int_{-\infty}^{z-x} dF_2(y) = \int_{-\infty}^{+\infty} F_2(z-x)\, dF_1(x) \\ &= \int_{-\infty}^{+\infty} F_2(z-x)\, f_1(x)\, dx, \end{aligned} \tag{6.16}$$

$$f(z) = \int_{-\infty}^{\infty} f_1(x)\, f_2(z-x)\, dx, \tag{6.17}$$

The rôles of F_1 and F_2 can, of course be interchanged (choose the simplest way!) and, as usual, we make the qualification that the expressions in terms of densities only hold when the latter exist.

The operations on the distributions which gives F in terms of F_1 and F_2, and f in terms of f_1 and f_2, are called *convolutions*. They are usually denoted by the symbols * and *, and we write $F = F_1 * F_2$, $f = f_1 * f_2$.

The operation can clearly be repeated to give the distribution of the sum of three independent random quantities (and so on for any finite sum). It follows from the definition that convolution is associative, commutative and even distributive. In the special case where all the summands are identically distributed (that is, have the same distribution function F), the convolution is denoted by F^{*n} (and f^{*n}).

The following is a brief summary of the other cases we considered:

$$\begin{aligned} \textit{product}: \qquad & F(z) = \int_{0}^{+\infty} F_2(z/x)\, dF_1(x), \\ & f(z) = \int_{-\infty}^{+\infty} \frac{1}{|x|} f_1(x)\, f_2(z/x)\, dx; \end{aligned} \tag{6.18}$$

$$\begin{aligned} \textit{quotient}: \qquad & F(z) = \int_{0}^{+\infty} F_2(zx)\, dF_1(x), \text{[38]} \\ & f(z) = \int_{-\infty}^{+\infty} |x| f_1(x)\, f_2(zx)\, dx; \end{aligned} \tag{6.19}$$

38 For the sake of brevity, the term $\int_{-\infty}^{0}$ (anti-symmetric) is omitted; if X is not certainly positive, it must be included.

$$\textit{distance}: \qquad F(z)=\int_{-z}^{z}\left[F_2\left(\sqrt{\left(z^2-x^2\right)}\right)-F_2\left(-\sqrt{\left(z^2-x^2\right)}\right)\right]\mathrm{d}F_1(x),$$

$$f(z)=\int_{-z}^{z}\frac{z}{\sqrt{\left(z^2-x^2\right)}}f_1(x)[f_2\left(\sqrt{\left(z^2-x^2\right)}\right) \tag{6.20}$$

$$+f_2\left(-\sqrt{\left(z^2-x^2\right)}\right)]\mathrm{d}x;$$

$$\textit{maximum}: \qquad F(z)=F_1(z)F_2(z), \qquad f(z)=F_1(z)f_2(z)+F_2(z)f_1(z). \tag{6.21}$$

6.9.7. *Synthetic characteristics for sums of independent random quantities.* Let Z be the sum of two or more independent random quantities; we shall include both $Z = X + Y$ and $Z = X_1 + X_2 + \ldots + X_n$ in order to draw attention both to the notationally simplest case and to the general one.

We shall consider now some of the points that can be made concerning their synthetic characteristics. We shall use the indices $i = 1, 2,\ldots, n$ for aspects concerning the summands, and $\overline{n}|$ for what concerns the sum of n terms; when the summands are identically distributed, we shall drop the indices.

In the case of the prevision, $m = \mathbf{P}(X)$, we have additivity (in all circumstances); for the variance, $\sigma^2 = \mathbf{P}(X - m)^2$, additivity holds when the summands are uncorrelated (and, *a fortiori*, when they are independent):

$$m_{\overline{n}|}=m_1+m_2+\ldots+m_n\left(=n.m\right) \tag{6.22}$$

$$\sigma^2_{\overline{n}|}=\sigma_1^2+\sigma_2^2+\ldots+\sigma_n^2 \quad \left(=n\sigma^2;\sigma_{\overline{n}|}=\sqrt{n}\sigma\right). \tag{6.23}$$

For the third-order moments, we have

$$\mathbf{P}\left(Z^3\right)=\mathbf{P}\left(X+Y\right)^3=\mathbf{P}\left(X^3\right)+3\mathbf{P}\left(X^2Y\right)+3\mathbf{P}\left(XY^2\right)+\mathbf{P}\left(Y^3\right),$$

and, in the case of independence,

$$\mathbf{P}\left(Z^3\right)=\mathbf{P}\left(X^3\right)+3\mathbf{P}\left(X^2\right)\mathbf{P}(Y)+3\mathbf{P}(X)\mathbf{P}\left(Y^2\right)+\mathbf{P}\left(Y^3\right).$$

For $Z = \Sigma X_i$, with the summands independently and identically distributed, if we denote by

$$M_1=m=\mathbf{P}(X),\ M_2=m^2+\sigma^2=\mathbf{P}\left(X^2\right),\ M_3=\mathbf{P}\left(X^3\right)$$

the moments (of 1st, 2nd and 3rd orders, respectively) of the summands, and by $(M_3)_{\overline{n}|}$ that of the sum, we have similarly

$$\left(M_3\right)_{\overline{n}|}=\sum_{ijh}\mathbf{P}\left(X_iX_jX_h\right)=nM_3+3n(n-1)M_1M_2+n(n-1)(n-2)M_1^3. \tag{6.24}$$

On the basis of this formula, the reader can see how things proceed in the general case by noting the following simple points (and these will not apply only to M_3, with summands not identically distributed, but to moments of any order, whether the summands are identically distributed or not):

the 3rd power (or the general *r*th power) of a sum of *n* terms is the sum of the n^3 (or n^r) products (including repetitions) of the summands three at a time (or *r* at a time);
the prevision of each product is (M_3) if it contains precisely the same factor X_i three times; $(M_2)_i(M_1)_j$ if the product is $X_iX_iX_j$; $(M_1)_i(M_1)_j(M_1)_k$ if the product is $X_iX_jX_k$ (with distinct factors); for *r* summands, things become more complicated, but the idea is the same;
in the case of identically distributed summands, it is sufficient to suppress the indices *i*, *j* and *k*, and count up the number of the three kinds of term M_3, M_2M_1, M_1^3 (and there are *n* choices for *i*; $3n(n-1)$ ways of putting a *j* in one of the three positions and an $i \neq j$ in the remaining two; $n(n-1)(n-2)$ ways of arranging the *n* elements three at a time); for a general *r*, we have products of the form $M_1^a M_2^b \ldots M_n^m$, with $a + 2b + 3c + \ldots + mn = n$, if the product contains *a* single factors, *b* which appear twice, *c* which appear three times,..., and *m* (either 0 or 1) *n*-tuples.

As far as the extreme values, inf *Z* and sup *Z*, are concerned, in the case of independence we can definitely say that inf $Z = \Sigma \inf X_i$ and sup $Z = \Sigma \sup X_i$ (in general one can only note the obvious inequalities, $\geqslant$ and $\leqslant$, respectively).

6.9.8. One obvious additional result is that for the sum of independent random quantities (i.e. the convolution of distributions) the range of variation of the distribution must increase: if $F = F_1 * F_2$,

$$\sup F - \inf F > \sup F_1 - \inf F_1;$$

(with equality only in the trivial case of F_2 concentrated at a single point).

The same conclusion holds, however, in a much more general context: the *dispersion* $l(p)$ must also increase (for all $0 < p \leqslant 1$; the above corresponds to the extreme case $p = 1$). Suppose that in the distribution *F* there is an interval of length *l* enclosing a mass $\geqslant p$; let the interval be $a, a + l$: if we assume that in F_1 every interval of length *l* contains a mass $< p$ (see 6.6.6) we are led to the following absurd conclusion:

$$p \leqslant F(a+l) - F(a)$$
$$= \int_{-\infty}^{\infty} \{F_1(a+l-x) - F_1(a-x)\} dF_2(x) < p \int_{-\infty}^{\infty} dF_2(x) = p.$$

It follows, as an important corollary, that, for the convolution, 'regularity' must increase: the resulting distribution enjoys all those regularity properties enjoyed by at least one of the component distributions. For example: the property of not having any masses greater than some given *p*; the property of continuity; or of being absolutely continuous; or of having a density never greater than some given bound; properties of existence or bounds for successive derivatives; or the property of being analytic.

It can easily be seen, for instance, that the mathematics used in 6.7.2 to construct a continuous distribution 'close' to some given one, was essentially an application of the following: given any random quantity, in order to obtain a distribution with density $\leqslant 1/\varepsilon$, it is sufficient to add to it a random quantity with a uniform distribution in the interval $[0, \varepsilon]$ (for example, a 'rounding error'). An 'accidental' error with a *normal* distribution – which we shall meet soon – is sufficient to make the distribution analytic.

In addition to the moments, $\gamma = \square^r$, which we have already considered, there is another class of previsions $F(\gamma)$ of great importance: that of the exponential functions $\gamma = a^{\square}$. The basic property of these functions yields, for $Z = X + Y$ (or $Z = \Sigma X_i$),

$$a^z = a^{X+Y} = a^X a^Y, \quad a^Z = a^{\Sigma_i X_i} = a^{X_1 X_2} \ldots a^{X_n},$$

so that, in the case of independence,

$$\mathbf{P}\left(a^Z\right) = \mathbf{P}\left(a^X\right)\mathbf{P}\left(a^Y\right), \mathbf{P}\left(a^Z\right) = \mathbf{P}\left(a^{X_1}\right)\mathbf{P}\left(a^{X_2}\right)\ldots\mathbf{P}\left(a^{X_n}\right). \tag{6.25}$$

We shall see shortly how this property can be exploited.

6.10 The Method of Characteristic Functions

6.10.1. The synthetic characteristics provide partial information of varying usefulness and interest; we have examined some of the most important kinds. One could ask, however, whether it is possible for a sufficiently rich set of 'synthetic characteristics' to be sufficient to completely characterize a distribution?

In terms of the $F(\gamma)$, the answer (in a general form) has already been given (in 6.4.4), since, in order to determine $F(\gamma)$, we said that it was sufficient to know $F(\gamma)$ for all continuous γ (it is also sufficient to know it for a subset which permits approximation to any desired degree of accuracy from above and below). It is known that in certain cases (for example, for bounded distributions) this can even be obtained by means of polynomials, and hence knowledge of (all) the moments, $F(\square^r)$, $r = 1, 2, \ldots, n, \ldots$, turns out to be sufficient (and, in fact, the researches of Tchebychev and others have dealt with this topic; Castelnuovo's treatise gives a masterly account of the research in this field). On the other hand, this method of moments also appears in the approach that we shall adopt.

This is the approach based on the property of the exponential function that we noted above. It consists in considering the prevision for the exponential function as the base varies in an appropriately chosen set (the reals, or, better still, complex values with absolute value =1). The method is called that of *generating functions*, or *characteristic functions* (according to the variant adopted). In order to avoid using more than one term (which is often misleading, since it prevents one seeing the essential identity of things expressed in slightly different forms) we shall always use the name 'characteristic function'.

This powerful technique has a rather curious history:[39] it has entered into consistent and systematic usage only recently (especially following the brilliant applications of it made by P. Lévy in about 1925), after having been discovered, applied, abandoned and then rediscovered in a variety of applications and circumstances (from De Moivre to Lagrange, from Laplace to Poisson).

6.10.2. In the simplest case (the original application of De Moivre), the method consists in noting that if X is a random *integer*, and t any real (or complex), then $\mathbf{P}(t^X) = \sum_h p_h t^h$ is a polynomial in which the coefficient of t^h is the probability of obtaining the value $X = h$ (h an integer, often – but not necessarily – positive). One also notes – and this is the *fundamental property* that we mentioned – that if X and Y are *stochastically independent* random quantities, so are t^X and t^Y, and hence

39 A clear, concise and essentially complete account can be found in H.L. Seal, 'The historical development of the use of generating functions in probability theory', *Bull. Ass. Actuaires Suisses*, **49** (1949), 209–228.

$$\mathbf{P}\left(t^{X+Y}\right)=\mathbf{P}\left(t^{X}t^{Y}\right)=\mathbf{P}\left(t^{X}\right)\mathbf{P}\left(t^{Y}\right). \tag{6.26}$$

If $\mathbf{P}(t^X) = \Sigma_h\, p_h t^h$ and $\mathbf{P}(t^Y) = \Sigma_k\, q_k t^k$, and we take the product

$$\sum_{hk} p_h q_k t^{h+k} = \sum_i t^i \sum_h p_h q_{i-h}, \tag{6.27}$$

we have an 'automatic' way of computing the probabilities

$$r_i = \mathbf{P}\left(X+Y=i\right)=\sum_h p_h q_{i-h}; \tag{6.28}$$

that is, of obtaining *the distribution of the sum*, $Z = X + Y$.

This fundamental property (that is, that the product $\mathbf{P}(t^X)\mathbf{P}(t^Y)$) corresponds to the sum $(X + Y)$ clearly holds even if X and Y are not integer, so long as t^X and t^Y continue to make sense. In order that this be so, one could limit oneself to t on the positive real axis, or, alternatively, write $t = \mathrm{e}^z$, with the convention that in place of $t^X = (\mathrm{e}^z)^X$ one considers $\mathrm{e}^{zX}(= \mathrm{e}^{(zX)})$, which always makes sense.[40]

Instead of $\mathbf{P}(t^X)$ we therefore consider $\mathbf{P}(\mathrm{e}^{zX})$ (which is equivalent when t is real and positive and z is real, and more general in that it allows the removal of these restrictions). If X has an unbounded distribution, $\mathbf{P}(\mathrm{e}^{zX})$ could diverge; this could never happen if z were purely imaginary (since then $|\mathrm{e}^{zX}| = 1$). In order to map the imaginary axis (which has this nice property we have just mentioned) onto the real axis (which is more convenient as the standard support for representing functions of a real variable) we set $z = \mathrm{i}u$, and then $t = \mathrm{e}^z = \mathrm{e}^{\mathrm{i}u}$; in this way $\mathbf{P}(\mathrm{e}^{\mathrm{i}uX})$ becomes a function of u, which is certainly defined for all u on the real axis (where, however, it will in general assume complex values), and possibly outside it as well.

But, in the general case, will knowledge of $\mathbf{P}(\mathrm{e}^{\mathrm{i}uX})$ be sufficient to determine the probability distribution? We shall see that the answer to this is yes. The answer is unconditional if we know $\mathbf{P}(\mathrm{e}^{\mathrm{i}uX})$ for all real u (or if we know $\mathbf{P}(\mathrm{e}^{zX})$ for all purely imaginary z); under suitable conditions, it also holds for $\mathbf{P}(t^X)$ and $\boldsymbol{P}(e^{zX})$ and for $t > 0$ and z real.

This justifies the name *characteristic function* given to

$$\phi(u)=\mathbf{P}\left(\mathrm{e}^{\mathrm{i}uX}\right) \tag{6.29}$$

(and sometimes also to $\mathbf{P}(\mathrm{e}^{zX})$); and the name *generating function* given to $g(t) = \mathbf{P}(t^X)$. We shall always use $\phi(u) = \mathbf{P}(\mathrm{e}^{\mathrm{i}uX})$, permitting ourselves to write (when X is an integer, and it is convenient to do so) $\phi(u) =$ (an expression in t) implying that $t \equiv \mathrm{e}^{\mathrm{i}u}$ (and we shall not speak of generating functions: one of the two terms is superfluous).

In the case of discrete distributions (masses p_h at the points x_h) or of distributions admitting a density function $f(x)$, the characteristic function can be expressed in the form

$$\phi(u)=\sum_h p_h \mathrm{e}^{\mathrm{i}ux_h} \tag{6.30′}$$

or

$$\phi(u)=\int \mathrm{e}^{\mathrm{i}ux} f(x)\mathrm{d}x, \tag{6.30″}$$

40 To the infinity of values $z = z_0 + 2k\mathrm{i}\pi$, having values e^z which coincide for a given t, there correspond different values for the nonexistent 't^x', i.e. $\mathrm{e}^{(z_0+2k\pi\mathrm{i})x}$.

respectively: in the general case, we have (using the Riemann–Stieltjes integral)

$$\phi(u) = \int e^{iux}\, dF(x) = F\left(e^{iu\square}\right)\left(\int = \int_{-\infty}^{+\infty}\right). \tag{6.30}$$

Of course, if one prefers to avoid the imaginary number under the prevision and integral signs, or if one wishes rather to give the real and imaginary parts separately, it suffices to recall that $e^{ix} = \cos x + i \sin x$ and to write

$$\phi(u) = \mathbf{P}(\cos uX) + i\mathbf{P}(\sin uX) = \int \cos ux\, dF(x) + i\int \sin ux\, dF(x). \tag{6.29′}$$

6.10.3. There is a one-to-one, onto and continuous correspondence between characteristic functions $\phi(u)$ and proper distributions $F(x)$. The inverse relation (in the simplest case, where

$$\int_{-\infty}^{\infty} |\phi(u)|\, du < \infty$$

and $f(x)$ is then continuous and bounded) is given by

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{*+\infty} e^{-iux}\phi(u)\, du^{41} \tag{6.31}$$

and has a symmetric relationship with equation 6.30″; this remarkable fact will be important in applications. By continuity we mean that *the convergence of* $\phi_n(u) \to \phi(u)$, *uniformly in any bounded interval, is equivalent to the convergence of* $F_n(x) \to F(x)$ *for all* x (*except for the discontinuity points of* F).

The fundamental property that we began with states that: *to the convolution of distributions,* $F = F_1 * F_2$ (or of densities $f = f_1 * f_2$) *there corresponds the product,* $\phi = \phi_1\phi_2$, *of characteristic functions*.

Moreover, to any linear combination, $F = \sum_h c_h F_h$, there corresponds the same linear combination, $\phi = \sum_h c_h \phi_h$. These properties in themselves are sufficient to solve many problems; they are also useful for deriving new distributions and for modifying distributions in order to make formulae like equation 6.31 applicable (by means of approximations) in cases where they are not directly applicable.

It is useful to bear in mind the following properties (for proofs and details see, for example, Feller, II, pp. 472 ff.): $\phi(u)$ is *continuous*; $\phi(0) = 1$ and $|\phi(u)| \leqslant 1$; the real part of $\phi(u)$ is *even* and the imaginary part *odd*; $\phi(u)$ is *real* if and only if the distribution is *symmetric*; changing X into aX, changes $F(x)$ into $F(x/a)$, and $\phi(u)$ into $\phi(au)$.

For the moments $\mathbf{P}(X^h) = M_h$ (where $\mathbf{P} = \check{\mathbf{P}}$!) which exist, the following expansion is valid

$$\phi(u) = 1 + iuM_1 - u^2 M_2/2! - iu^3 M_3/3! + \ldots + (iu)^h M_h/h! + \ldots$$

---

41 The asterisk at the upper limit of the integral sign means that the principal value (in the Cauchy sense) is to be understood: i.e. $\lim \int_{-a}^{a}$ as $a \to \infty$.

Formula 6.31 is the classical Fourier *inversion theorem*.

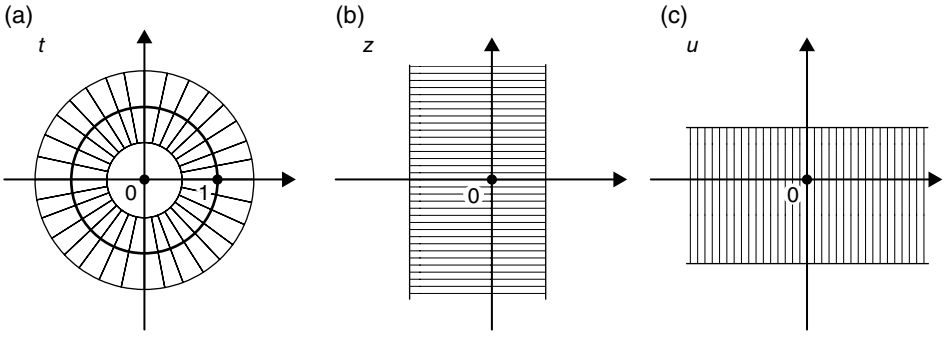

Figure 6.6 The planes of the three variables *t*, *z* and *u*, in terms of which the characteristic function can be expressed, together with the lines or regions where it is defined. Usually we operate in terms of *u* (the Fourier transform); $z = iu$ and $t = e^z = e^{iu}$ are occasionally to be preferred (the Laplace and Mellin transforms, respectively).

(and corresponds, formally, to $\mathbf{P}(e^{iuX}) = \mathbf{P}(1 + iuX - u^2X^2/2!-\ldots)$). If all the moments exist, the series has a nonzero radius of convergence ρ, and $\phi(u)$, and therefore the distribution is completely determined by the sequence of moments.[42]

These and other properties reveal a relationship to be borne in mind in the following qualitative sense: the smaller the 'tails' of the distribution at infinity (i.e. the faster $F(x)$ tends to 0 or 1), the more regular the behaviour of $\phi(u)$ near the origin; the smoother (in terms of differentiability etc.) the distribution is, the more regular is the behaviour of $\phi(u)$ at infinity.

6.10.4. *Geometrical representation of the mathematical nature of the problems.* We note that the functions of *u*, *z* and *t* that we have been considering are the transformations of the distribution function known in analysis as the Fourier, Laplace and Mellin transforms, respectively. As we have already indicated (but reiterate for the sake of anyone who has come across these transforms separately and has not noticed the fact), we are always dealing with precisely the same transform, apart from a change of variable. Figure 6.6 indicates, schematically, the planes of the (complex) variables *u*, $z = iu$ and $t = e^z$; the line on which the function is always defined is marked in heavily (the real axis in the case of *u*, the imaginary axis for *z*, and the unit circle for *t*), and the striped region indicates where it is defined in the analytic case:

$$-\alpha' < \mathscr{I}(u) = \mathscr{R}(z) < \alpha'', \; |t'| < |t| < |t''|^{43}$$

(where $0 \leqslant \alpha', \alpha'' \leqslant +\infty; 0 \leqslant |t'| = e^{\alpha'} \leqslant 1 \leqslant e^{\alpha''} = |t''| \leqslant \infty$).

---

42 Since $1/\rho = \limsup (e/n) \sqrt[n]{|M_n|}$, a necessary and sufficient condition for the function to be analytic is that $\sqrt[n]{|M_n|}$, the mean of order *n*, does not increase faster than *n* (i.e. remains $\leqslant Kn$ with *K* finite). The necessary and sufficient condition for the distribution to be determined by the moments is that the sum of the reciprocals, $1/\sqrt[n]{|M_n|}$, diverges (Carleman). This is a little less restrictive than the above, which implies that the sum of the reciprocals, $\geqslant 1/Kn = K^{-1}(1/n)$, diverges almost as rapidly as the harmonic series.

43 The annulus of convergence for the Laurent series (Figure 6.6a); the strips for the Dirichlet series (Figure 6.6b and, changing axes, Figure 6.6c). $\mathscr{R}$ and $\mathscr{I}$ denote the real and imaginary parts.

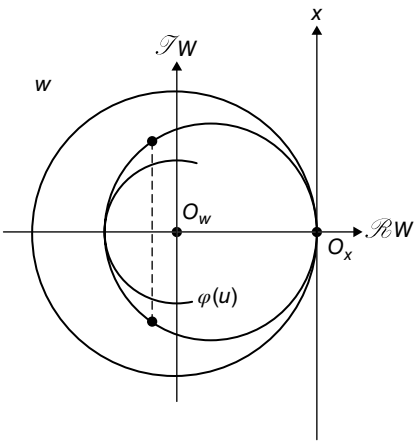

Figure 6.7 The plane of $w = \phi(u)$, and the interpretation of $\phi(u)$ as the barycentre of the distribution 'wrapped around the circumference $|w| = 1$'.

We have so far seen illustrations of the complex planes of the three variables (t, z, u). In order to 'visualize' the meaning and the properties of the characteristic function $\phi(u)$ (for u real) in the complex plane of $w = \phi(u)$, we draw it (Figure 6.7) indicating the unit circle, $|w| = 1$, and the tangent at the point $w = 1$ (the straight line $\mathscr{R}(w) = 1$). This point is denoted by 0_x, because it is the origin of the x coordinate, thought of both as the abscissa on this tangent line and as parameter (angle or arc length) on the circumference. In order to avoid confusion, the origin $w = 0$ has been denoted by 0_w.

If we think of the distribution of X as located on the x-axis, then e^{iX} has the same distribution 'wrapped around the unit circle', and similarly for u^X and e^{iuX} (with only a modification of scale from 1 to u, reflected if u is negative). The characteristic function $\phi(u) = \mathbf{P}(e^{iuX})$ is the barycentre (necessarily inside the circle, unless the distribution is concentrated at a single point), and it follows therefore that $|\phi(u)| \leqslant 1$. If $u = 0$, we always have, of course, $\phi(u) = 1$; in general, however, we have $|\phi(u)| < 1$, the only other exceptional cases being the following. Firstly, a trivial case consisting of a single mass concentrated at $x = a$; in this case we always have $\phi(u) = e^{iua}$, and, hence, $|\phi(u)| = 1$. The second exception is that of a distribution concentrated at the points of an arithmetic progression, $x = c \pm 2k\pi/u_0$; clearly $|\phi(u_0)| = 1$, and the same will hold for all multiples of u_0.

If we think in terms of the graph of $w = \phi(u)$, many properties (those we have already mentioned and others) become obvious. As an example, the change from X to $-X$ implies that the distribution (on the line, or wrapped around the circle) is reflected in the real axis; the same is also true for the barycentre, so that the characteristic function of $-X$ is the conjugate of the $\phi(u)$ corresponding to X; $\phi(-u) = \phi^*(u)$. An important corollary follows: given any $\phi(u)$, we can obtain a symmetric characteristic function, $|\phi(u)|^2 = \phi(u)\phi^*(u)$. The corresponding distribution is called the *symmetrized*[44] version of the $F(x)$ we started with, and is obtained from the convolution of $F(x)$ and $1 - F(-x)$;

44 Another form of symmetric distribution is obtained by taking the average of the given distribution $F(x)$ and its reflection $1 - F(-x)$; this gives a distribution function $\frac{1}{2}[1 + F(x) - F(-x)]$ with characteristic function $\frac{1}{2}[\phi(u) + \phi(-u)]$. It is the distribution we obtain when we toss a coin before deciding whether to take $+|X|$ or $-|X|$.

it is the distribution of the difference $X_1 - X_2$, where X_1 and X_2 are independent, both with distribution $F(x)$.[45]

For u purely imaginary (and we shall write $u = \mathrm{i}\nu$, with real ν, so that $\nu = \mathrm{i}u = z$), we have, separating the contribution of the probability distribution on the negative semi-axis from that on the positive axis, and from that concentrated at the origin, if any ($p_0 = F(+0) - F(-0)$),

$$\phi(-\mathrm{i}\nu) = \int_{-\infty}^{\infty} \mathrm{e}^{\nu x}\,\mathrm{d}F(x) = \int_{-\infty}^{0} \mathrm{e}^{\nu x}\,\mathrm{d}F(x) + \int_{0}^{\infty} \mathrm{e}^{\nu x}\,\mathrm{d}F(x) + p_0. \tag{6.32}$$

The contribution in $[-\infty, 0]$ is clearly finite for $\nu \geqslant 0$, and possibly for negative ν between 0 and some $-\alpha'$ (everywhere if $\alpha' = \infty$); by symmetry, the contribution in $[0, \infty]$ is finite for $\nu \leqslant 0$, and possibly for positive ν between 0 and some α'' (everywhere if $\alpha'' = \infty$). If it exists in the interval $[-\alpha', \alpha'']$ of the imaginary axis, ϕ is positive, real and concave (upwards), like each of the $\mathrm{e}^{\nu x}$ of which it is a mixture. The meaning of the bounds, $-\alpha'$ and α'', and some other aspects, becomes clear if we introduce the notion of *twinned*[46] distributions, a notion which is of interest in its own right.

The twins of $F(x)$ (and the relationship is mutual) are defined to be those $F_\nu(x)$ for which

$$\mathrm{d}F_\nu(x) = K\mathrm{e}^{\nu x}\,\mathrm{d}F(x), \quad \text{with } 1/K = \phi(-\mathrm{i}\nu); \tag{6.33}$$

this defines distributions whenever $\phi(-\mathrm{i}\nu)$ makes sense.

When the densities exist, we have

$$f_\nu(x) = K\mathrm{e}^{\nu x} f(x), \tag{6.34}$$

and the meaning may be clearer (because the notation is more familiar). We see immediately that the characteristic function of $F_\nu(x)$ is given by $\phi_\nu(u) = K\phi(u + \mathrm{i}\nu)$ (where $\phi = \phi_0$ is the characteristic function of $F(x)$), and it follows that $\phi(u)$ is defined throughout the strip $-\alpha' < \mathscr{I}(u) < \alpha''$ (in other words, there is no further restriction due to singularities outside the imaginary axis for u; in particular, if α' and α'' are both positive, $\phi(u)$ is analytic, and the minimum of the two bounds is the radius of convergence).

Expressed in a nonmathematical way, the conclusion is that $\phi(\mathrm{i}\nu)$ exists (and hence so does $\phi(u)$ over the entire line $\mathscr{I}(u) = \nu$) if the twin distribution $F_\nu(x)$ exists, and that this happens if the tail of $F(x)$ on the positive semi-axis (for positive ν; conversely for negative ν) is thinner than the tail of the exponential distribution $f(x) = K\mathrm{e}^{-\nu x}$; α' and α'' are zero, infinite or finite, depending on whether the tail (on the left or on the right) is fatter or thinner than every exponential, or comparable with an exponential, respectively.

---

45 Symmetrized distributions are also considered in the statistical context. The prevision of $X_1 - X_2$ is zero, but the quadratic prevision and that of $|X_1 - X_2|$ constitute 'indices of variability' (the first one is clearly simply $\sigma(X)$ multiplied by $\sqrt{2}$); $\mathbf{P}(|X_1 - X_2|)$ turns out to be the concentration ratio multiplied by $2\mathbf{P}(X)$, which, for a given $\mathbf{P}(X)$, is the maximum possible value: see 6.6.3.

46 The term *conjugate* (see Keilson, 1965) is used in other contexts (see, for example Chapter 12, 12.4.2). I therefore suggest the term given in the text. Feller (II, p. 410) refers to the property in question as the *translation principle* (but, as far as I know, does not give a name to such distributions).

6.11 Some Examples of Characteristic Functions

6.11.1. This is a convenient point at which to note and calculate explicitly the characteristic functions of some common distributions. In part, these will be cases of importance for applications; in part, they will be examples whose main purpose is to show how one can often avoid direct calculation with shrewd use of the properties of characteristic functions, keeping an eye on their interpretations. Until we actually illustrate these ideas with reference to the applications, the sense of this must inevitably remain somewhat unclear, but just a brief mention of the nature of the applications will suffice to give the basic idea.

6.11.2. In the case of an event E (with probability $p = \mathbf{P}(E)$), or for a bet $s(E - p^*)$ on E (with gain s if E occurs, loss p^*s if it does not – the bet is fair if $p^* = p$), we have, respectively,

$$\phi(u) = \mathbf{P}(e^{iuE}) = \tilde{p}e^{iu0} + pe^{iu1} = 1 + p(e^{iu} - 1), \tag{6.35}$$

$$\begin{aligned}\phi(u) &= \mathbf{P}(e^{ius(E-p^*)}) = e^{-iusp^*}\mathbf{P}(e^{iusE}) = e^{-iusp^*}[1 + p(e^{ius} - 1)] \\ &= (1-p)e^{-iusp^*} + pe^{ius(1-p^*)}\end{aligned} \tag{6.36}$$

(here, and elsewhere, it is sufficient to apply the property relating to additive and multiplicative constants: $\mathbf{P}(e^{iu(cX+k)}) = e^{iuk}\mathbf{P}(e^{icuX})$, in other words, change $\phi(u)$ into $e^{iuk}\phi(cu)$).

In the particular case where $s = 2$, $p = p^* = \frac{1}{2}$, we have a fair bet at the game of Heads and Tails with gains ±1: the above reduces to

$$\phi(u) = \frac{1}{2}\left(e^{iu} + e^{-iu}\right) = \cos u, \tag{6.37}$$

whereas

$$\mathbf{P}\left(e^{iuE}\right) = \frac{1}{2}\left(1 + e^{iu}\right). \tag{6.37′}$$

In the case of n independent tosses, the gain $2Y - n$, and the number of successes $Y = E_1 + E_2 + \ldots + E_n$, have characteristic functions

$$\phi(u) = \cos^n u, \tag{6.38}$$

and

$$\phi(u) = \left[\frac{1}{2}\left(1 + e^{iu}\right)\right]^n, \tag{6.38′}$$

respectively (the sum of independent random quantities = convolution = product of characteristic functions; in particular, this becomes a power if the distributions are identical).

Similarly, if p is now taken to be general (and we continue to assume stochastic independence), the number of successes, Y has the characteristic function

$$\phi(u) = \left[1 + p\left(e^{iu} - 1\right)\right]^n. \tag{6.38″}$$

This is the so-called Bernoulli distribution: the limit-case, obtained by letting n tend to ∞ with the prevision $np = a$ held constant, is called the Poisson distribution. This gives

$p_h = e^{-a}a^h/h!$, and hence its characteristic function is given by

$$\phi(u) = \lim\left[1 + \frac{a}{n}\left(e^{iu} - 1\right)\right]^n = e^{a\left(e^{iu}-1\right)}. \tag{6.39}$$

In all cases where the possible values are non-negative integers (like the above, relating to Y, 'the number of successes') the characteristic function is a polynomial (or a power series) in $t = e^{iu}$ with coefficients $p_h = \mathbf{P}(Y = h)$:

$$\phi(u) = \sum_h p_h t^h = \sum_h p_h e^{iuh}.$$

Knowing this, we could have obtained equations 6.38′, 6.38″ and 6.39 directly from the knowledge of the p_h; conversely, to find the latter from the characteristic function we expand in powers of e^{iu}.

Let us have another look at three distributions of this type (having integer values); we consider the *uniform*, *geometric and logarithmic*.

For the *uniform* distribution ($p_h = 1/n$; $1 \leqslant h \leqslant n$) one has

$$\begin{aligned}\phi(u) &= 1/n\sum_{h=1}^{n} e^{iuh} = \frac{e^{iu(n+1)} - e^{iu}}{n\left(e^{iu} - 1\right)} \\ &= \frac{1}{n}e^{iu\frac{1}{2}(n+1)}\frac{e^{\frac{1}{2}iun} - e^{-\frac{1}{2}iun}}{e^{\frac{1}{2}iu} - e^{-\frac{1}{2}iu}} = e^{iu\frac{1}{2}(n+1)}\frac{\sin\frac{1}{2}nu}{n\sin\frac{1}{2}u}.\end{aligned} \tag{6.40}$$

For the *geometric* distribution ($p_h = Kq^h$, $0 < q < 1$, $K = 1 - q$; $0 \leqslant h < \infty$) one has

$$\phi(u) = (1-q)\sum_{h=0}^{\infty} q^h e^{iuh} = K/\left(1 - qe^{iu}\right) = (1-q)/\left(1 - qe^{iu}\right). \tag{6.41}$$

For the *logarithmic* distribution ($p_h = Kq^h/h$, $0 < q < 1$, $K = -\log(1 - q)$; $1 \leqslant h < \infty$) one has

$$\phi(u) = K\sum_{h=1}^{\infty} q^h e^{iuh}/h = -K\log\left(1 - qe^{iu}\right) = \log\left(1 - qe^{iu}\right)/\log(1-q). \tag{6.42}$$

6.11.3. Let us now turn our attention to some continuous distributions: we shall present the density functions $f(x)$ and the characteristic functions $\phi(u)$, always expressed in the most convenient standard form (since any transformation from X to $cX + k$ can be easily dealt with).

The *normal* distribution (sometimes known as the 'error' distribution) will be well known to everyone, although we have not yet dealt with it explicitly. We shall give a more extensive treatment in Chapter 7 (Section 7.6).

The *standardized*, or normalized, distribution, with prevision = 0 and variance = 1, has density and characteristic function given by[47]

47 That this is the value of the normalization constant is well known from analysis. We shall, in any case, prove this (Chapter 7, 7.6.7) at a more appropriate and meaningful time.

$$f(x) = K\mathrm{e}^{-\frac{1}{2}x^2} \left(K = \frac{1}{\sqrt{2\pi}} \right), \tag{6.43}$$

$$\phi(u) = \mathrm{e}^{-\frac{1}{2}u^2}. \tag{6.44}$$

Direct calculation is straightforward (if we operate in the complex field, using the substitution $y = x - iu$; a little less straightforward if we proceed differently, or if we do not assume the form we want).

A convolution of normal distributions is also a normal distribution (in other words, the sum of independent random quantities having normal distributions also has a normal distribution). We express this fact by saying that the normal distribution is *stable*. In fact, one has $\mathrm{e}^{-\frac{1}{2}(au)^2}\,\mathrm{e}^{-\frac{1}{2}(bu)^2} = \mathrm{e}^{-\frac{1}{2}(cu)^2}$; in other words.

$$\phi(au)\phi(bu) = \phi(cu), \text{ where } c = \sqrt{\left(a^2 + b^2\right)}. \tag{6.45}$$

The scale parameters (a, b, c) are, in fact, the standard deviations, so it follows that the composition should take place according to Pythagoras' theorem (as is always the case for a finite sum). Observe also that

$$\phi^n(u) = \phi\left(\sqrt{n}.u\right) \tag{6.46}$$

and that

$$\phi^t(u) = \phi\left(\sqrt{t}.u\right) \tag{6.46'}$$

for any positive real t (and not only for integer n). The fact that $\phi^t(u)$ is always a characteristic function means that the distribution is *infinitely divisible* (e.g. into $(\phi^{1/n}(u))^n$). We have already encountered another example of an infinitely divisible distribution (although we did not point it out at the time), the Poisson, whose characteristic function (equation 6.39), contains an arbitrary constant as exponent (in equation 6.39 it was denoted by a). We shall soon come across other examples; the general form of infinitely divisible distributions, and the subclass of the stable distributions, will be given in Chapter 8, along with some of the important properties.

The *uniform* distribution (taken over [−1, +1]) has

$$f(x) = \frac{1}{2}\left(|x| \leqslant 1\right), \tag{6.47}$$

$$\phi(u) = \frac{\sin u}{u}. \tag{6.48}$$

The calculation is straightforward (it can also be obtained from the discrete case, equation 6.40), by letting $n \to \infty$ with nu = constant, along with obvious changes of origin and scale).

For the sum of two (independent) random quantities having this distribution we obtain

$$f(x) = \frac{1}{2}\left(1 - \frac{1}{2}|x|\right)\left(|x| \leqslant 2\right), \tag{6.49}$$

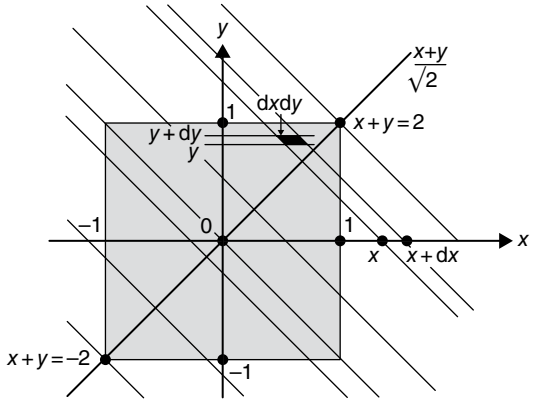

Figure 6.8 The convolution of uniform distribution.

$$\phi(u) = \left(\sin u\right)^2 / u^2; \tag{6.50}$$

which is called the 'triangular distribution', on account of the form of the graph of the density function (this could be deduced from the definition without any need for calculations: it is the orthogonal projection onto the diagonal of a square of a mass uniformly distributed on it; Figure 6.8).

The characteristic function is positive: it follows immediately, therefore, that, conversely, there is a distribution (on $-\infty \leqslant x \leqslant +\infty$) with density and characteristic function given by

$$f(x) = \frac{1}{\pi}\frac{\left(\sin x\right)^2}{x^2}, \tag{6.51}$$

$$\phi(u) = \left(1 - \frac{1}{2}|u|\right)\left(|u| \leqslant 2\right). \tag{6.52}$$

This distribution is not, in itself, very interesting. It is, however, of great importance in that one can immediately deduce from it conclusions of some generality. By means of mixtures of triangular distributions (on different ranges) we can obtain any distribution whose density has a polygonal graph (symmetric with respect to the origin, decreasing and concave upwards on either side of the origin). In the limit, we can obtain any curve with these kind of properties. By inversion, any function having such behaviour *is a characteristic function*: this is Pólya's criterion. The fact that in this way we can obtain characteristic functions which are zero outside a finite interval is also of some importance (Figure 6.9).

In a similar way, we obtain $\phi(u) = (\sin u)^n/u^n$ as the characteristic function of the sum of n independent random quantities which are uniformly distributed in $[-1, +1]$; this corresponds to the density of the projection onto the diagonal of an n-dimensional cube of a mass uniformly distributed on it, and is represented by polynomials of degree $n - 1$ which vary on each of the n intervals of length 2 into which the interval $[-n, +n]$ is divided by the projections of the vertices of the cube. Think of the ordinary cube, $n = 3$ (the areas of the sections are first triangular, then hexagonal, then triangular again).

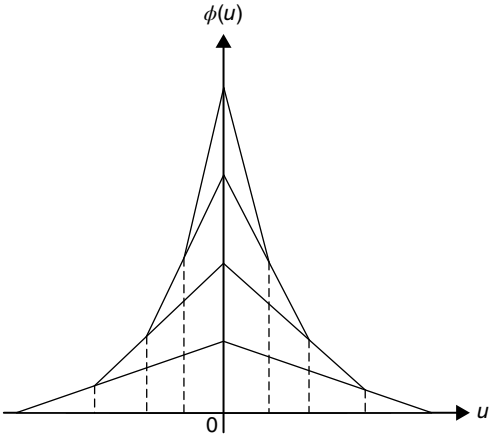

Figure 6.9 Characteristic functions constructed on the basis of Pólya's argument.

The inversion (as for $n = 2$) can be made for any even n (since then the characteristic function has to be positive).

For the *exponential distribution*,

$$f(x) = e^{-x} \quad (x \geqslant 0) \tag{6.53}$$

and

$$\phi(u) = 1/(1 - iu); \tag{6.54}$$

this is a special case ($t = 1$) of the *gamma* distribution, defined by

$$\begin{aligned} &f(x) = Kx^{t-1}e^{-x} \quad (x \geqslant 0) \quad \text{with} \quad K = 1/\Gamma(t), \\ &\Gamma(t) = \int_0^\infty x^{t-1}e^{-x}dx = (t-1)! \ \text{for integer} \, t, t > 0, \end{aligned} \tag{6.55}$$

$$\phi(u) = 1/(1 - iu)^t. \tag{6.56}$$

The fact that t appears as an exponent in $\phi(u)$ (or, more precisely, in $\phi^t(u)$) implies that these distributions have the property of being infinitely divisible.

By symmetrization of the gamma distribution, we obtain distributions whose characteristic functions are given by

$$\phi(u) = \left[1/(1 - iu)^t\right]\left[1/(1 - iu)^t\right]^* = 1/\left(1 + u^2\right)^t \tag{6.57}$$

(and these are also infinitely divisible). In particular, for $t = 1$, we have the two-sided exponential distribution, with

$$f(x) = \frac{1}{2}e^{-|x|}, \tag{6.58}$$

$$\phi(u) = 1/\left(1 + u^2\right). \tag{6.59}$$

By inversion, we obtain

$$f(x) = 1/\left[\pi\left(1+x^2\right)\right], \tag{6.60}$$

$$\phi(u) = \mathrm{e}^{-|u|}; \tag{6.61}$$

the Cauchy distribution. This is infinitely divisible (for $t > 0$, $(\mathrm{e}^{-|u|})^t = \mathrm{e}^{-|tu|}$ is the characteristic function of $f(x) = K/(1 + (x/t)^2)$) and, since $f(x)$ remains invariant (apart from changes of scale), the distribution is also *stable* (like the normal). Its invariance is infinite, as can be seen directly (from $f(x)$ being of second-order smallness) or from the irregularity of $\phi(u)$ at the origin.

6.11.4. Knowing the characteristic functions of certain distributions enables us – using products, powers, conjugation, linear combinations, limits and so on – to obtain innumerable others, as required for various applications, and corresponding to distributions whose densities cannot in many cases be expressed explicitly.

Let us examine some of the more interesting examples of mixtures; those given by the sum of N independent, identically distributed random quantities X_h when N itself is also random. If at each step there is a probability p of stopping and $q = 1 - p$ of continuing, the N has a geometric distribution; that is,

$$p_n = \mathbf{P}(N = n) = Kq^n \ \left(K = (1-q)\right).$$

If it turned out that $N = n$, the characteristic function of the sum would be $\chi^n(u)$, where $\chi(u)$ denotes the characteristic function of each X_h; the characteristic function of the unconditional sum is hence given by the mixture

$$\phi(u) = \sum_{n=0}^{\infty} Kq^n \chi^n(u) = K/\left[1 - q\chi(u)\right] = (1-q)/\left[1 - q\chi(u)\right]. \tag{6.62}$$

Formally, it is sufficient to substitute in equation 6.41, replacing the characteristic function $e^{\mathrm{i}u}$ of each of the summands '1' by the characteristic function $\chi(u)$ of X_h. Following the same rule in the general case, one obtains.

$$\phi(u) = \sum_n p_n \chi^n(u), \tag{6.63}$$

and, in the particular cases of N having the Bernoulli or Poisson distribution, we have (see equations 6.38″ and 6.39)

$$\phi(u) = \left[1 + p\left(\chi(u) - 1\right)\right]^m \tag{6.64}$$

and

$$\phi(u) = \mathrm{e}^{a\left[\chi(u)-1\right]}, \tag{6.65}$$

respectively. In equation 6.64 we used m in the exponent rather than n (which is now used to denote particular values of N); the interpretation (for example in the case of a game) is as follows: an individual has the right to m trials, each having probability p of success; he then has n successes ($0 \leqslant n \leqslant m$), and receives a random prize X_h for each success. The Poisson case can, for the present, be regarded as a limit-case (but will be seen to have a much more interesting interpretation when viewed as a 'random process'; see Chapter 8).

When a characteristic function $\chi(u)$ is infinitely divisible, that is, $\chi^t(u)$ is a characteristic function for any $t > 0$ (not only for t integer), one need not limit oneself to mixtures involving integer powers (equation 6.61), but can also consider sums of the form

$$\phi(u) = \sum_n p_n \chi^{t_n}(u), \quad \text{for any } t_n > 0, \tag{6.66}$$

or even

$$\phi(u) = \int_0^\infty p(t)\chi^t(u)\,\mathrm{d}t \left(\text{with } p(t) \geqslant 0, \int_0^\infty p(t)\mathrm{d}t = 1 \right). \tag{6.67}$$

6.11.5. If we take a random quantity X and add on a random quantity Δ, which is small and has appropriate regularity properties, then $X + \Delta$ will differ only slightly from X (it is as though we intentionally measure X with a small error), but will enjoy the regularity properties possessed by Δ (and perhaps some others as well). As we shall see, this can turn out to be very useful.

For example, suppose Δ is chosen to have a uniform distribution between $\pm\delta$, with density $1/2\delta$. In this case, $X + \Delta$ will always have a density $\leqslant 1/2\delta$, whatever the distribution of X (see 6.9.8). If we take a triangular distribution for $\Delta(f(x) = K(1 - |x|/\delta): K = 1/\delta)$, $X + \Delta$ will have a density which is $\leqslant 1/\delta$ everywhere, and the derivatives of the density will also be $\leqslant 1/\delta^2$ (in absolute value). Similar bounds obtain when Δ is taken to be normal ($m = 0$, $\sigma = \delta$).

In terms of characteristic functions, this results in $\phi(u)$, the characteristic function of X, being multiplied by the characteristic function of Δ; in the cases mentioned above, we consider

$$\phi(u)\left(\sin\delta u\right)/\delta u, \quad \phi(u)\left(\sin\delta u\right)^2/(\delta u)^2, \phi(u)\mathrm{e}^{-\frac{1}{2}(\delta u)^2}.$$

This device often enables us to reduce problems posed in terms of general distributions to a framework in which suitable regularity conditions are obeyed.

In particular, we observe that if Δ is assumed to have the first form mentioned above (uniform over $\pm\delta$ then $f_\delta(x)$ the density of $X + \Delta$, is precisely the *average density* of X in the interval $x \pm \delta$; in other words,

$$f_\delta(x) = \left[F(x+\delta) - F(x-\delta)\right]/2\delta.^{48}$$

Informally, this formula says the following: the probability of $X + \Delta$ lying between $x \pm \frac{1}{2}\mathrm{d}x$ is the probability (of the necessary condition) that X lies inside $x \pm \delta$, since, conditional on $X = x_0$ (x_0 any point in $x \pm \delta$) the density of $X + \Delta$ at x is always the same, $1/2\delta$. More formally, considering the convolution for $X + \Delta$ (see 6.9.6), we have

$$\begin{aligned} f_\delta(x) &= \int f(x).(1/2\delta)\left(|z-x| \leqslant \delta\right)\mathrm{d}x \\ &= (1/2\delta)\int_{z-\delta}^{z+\delta} f(x)\mathrm{d}x = 1/2\delta\left[F(z+\delta) - F(z-\delta)\right] \end{aligned}$$

48 The fact that $f_\delta(x)$ is discontinuous and undefined at those points (at most a countable number) at distance δ (to the left or right) from discontinuity points of $F(x)$ (points with concentrated mass for X) is irrelevant.

(clearly, we could have considered the general case straightaway, by writing $\mathrm{d}F(x)$ instead of $f(x)\,\mathrm{d}x$). This formula may be used to obtain $F(x'') - F(x')$ for a preassigned interval (x', x''). In fact, it suffices to put $z = \frac{1}{2}(x' + x'')$, $\delta = \frac{1}{2}(x'' - x')$. In particular, to obtain $F(x) - F(0)$, it is enough to put $z = \delta = \frac{1}{2}x$. We have, therefore,

$$F(x'') - F(x') = (x'' - x') f_{\frac{1}{2}(x''-x')}\left(\frac{1}{2}(x' + x'')\right), \quad F(x) - F(0) = x f_{\frac{1}{2}x}\left(\frac{1}{2}x\right).$$

The characteristic function of $f_\delta(x)$ is given by $\phi(u)(\sin \delta u)/\delta u$, so that we obtain the following inversion formula for passing from the characteristic function $\phi(u)$ to the distribution function $F(x)$:

$$F(x) - F(0) = \frac{1}{2\pi} \int_{-\infty}^{*+\infty} \mathrm{e}^{-\frac{1}{2}\mathrm{i}ux} \phi(u) \frac{\sin \frac{1}{2}ux}{\frac{1}{2}ux} \mathrm{d}u = \int_{-\infty}^{*+\infty} \phi(u) \frac{\mathrm{e}^{\mathrm{i}ux} - 1}{\mathrm{i}u} \mathrm{d}u. \tag{6.68}$$

This (or one of the alternative forms) is the standard result, usually proved on the basis of the Dirichlet integral; this is a more laborious method and, in the words of Feller (II, p. 484), 'detracts from the logical structure of the theory'.

6.11.6. A more intuitive and expressive way of interpreting and explaining this is as follows: we think of the characteristic function – and let us take the simplest case, $\phi(u) = \sum p_h \mathrm{e}^{\mathrm{i}ux_h}$ – as a mixture of oscillations of various frequencies x_h and intensities p_h (the variable u being thought of as time). The formula for determining the components of the mixture given $\phi(u)$ (or, if we think in terms of light, for separating it into its monochromatic components), corresponds to a device capable of filtering out lines or bands. In order to discover whether a component of frequency x_0 exists, and in this case to isolate it and determine its intensity p_0, we must have a monochromatic filter. This is precisely what is achieved by the operation of computing the mean value (over a long period) of $\phi(u)$ multiplied by $\mathrm{e}^{-\mathrm{i}ux_0}$; in more precise terms, the operation of computing

$$\lim_{a\to\infty} \frac{1}{2a} \int_{-a}^{a} \phi(u) \mathrm{e}^{-\mathrm{i}ux_0} \mathrm{d}u = \lim_{a\to\infty} \sum_h p_h \frac{1}{2a} \int \mathrm{e}^{\mathrm{i}u(x_h - x_0)} \mathrm{d}u. \tag{6.69}$$

We see immediately, however, that if $x_0 \neq x_h$ the mean value (over any period, and hence asymptotically, on any very long interval $[-a, a]$) is zero. Only if x_0 coincides with one of the x_h does the integrand reduce to $e^{\mathrm{i}u0} \equiv 1$, the mean value to 1, and the result to p_0 (that is, the p_h for which x_h is our x_0).

The other operations can be regarded as band filters, used to obtain the sum of the p_h corresponding to frequencies x_h contained in some given interval $[x', x'']$ and so on.

6.12 Some Remarks Concerning the Divisibility of Distributions

A distribution obtained by the convolution of others is said to be divisible into the latter (its factors); G is a factor of F if we can write $F = G * H$ (for some suitable H). In terms of characteristic functions, this means that $\phi(u)$ can be expressed as a product of functions $\phi_h(u)$, each of which is also a characteristic function.

We have already seen the example of infinitely divisible distributions that can be defined (in the simplest, but also the most meaningful way) as those for which $\phi(u) = [\phi(u)^{1/n}]^n$ for any n (that is, $\phi(u)^{1/n}$ is a characteristic function for every n). Although we shall have no reason to give a systematic treatment of this topic, we shall, from time to time, come across problems where divisibility enters in. For this reason, it is appropriate at this stage to briefly mention it, for the sole purpose of warning against the errors that can arise if one proceeds by analogy with factorization as it occurs in arithmetic or algebra. Anyone interested in pursuing the subject more deeply should read P. Lévy (1937), pp. 190–195, and the references cited there, or the recent survey by M. Fisz, in *Ann. Math. Stat.* (1962), 68–84; the latter contains a bibliography.

There exist distributions that are not divisible: for example, it is clear that a distribution with only four possible values, 0, a, b and c (in increasing order) cannot be divisible unless $c = a + b$ (in which case we must have Z = $X + Y$, with 0 and a the possible values for X, 0 and b the possible values for Y); given this fact, we see that the four probabilities p_0, p_a, p_b and p_c cannot be chosen arbitrarily (subject only to their sum =1), because they must be of the form $p_0 = (1 - \alpha)(1 - \beta)$, $p_a = \alpha(1 - \beta)$, $p_b = (1 - \alpha)\beta$ and $p_c = \alpha\beta$ (where $\alpha = \mathbf{P}(X = a)$ and $\beta = \mathbf{P}(X = b)$; an extra condition must hold, leaving two degrees of freedom instead of three).

In general, there are no uniqueness type properties for factorizations; a distribution always admits a decomposition into an infinitely divisible distribution and indivisible ones (Khintchin's theorem); there may be infinitely many of the latter, or none; or it may be that the former is not present. We can also have, in general, various, different factorizations, combining different factors, without even a sharp dividing line between the factor which is infinitely divisible and the others. In fact, it can happen that an infinitely divisible distribution turns out to be a product of indivisible factors when factorized in a particular way.

There are, however, important cases in which the factorization is unique, and, in fact, reduces to the *trivial* factorization – the decomposition into factors $[\phi(u)]^{t_h}$ (with $t_h > 0$, $\sum t_h = 1$) with $\phi(u)$ infinitely divisible. This is the case for the normal distribution (so that, if $X + Y = Z$ has a normal distribution and X and Y are independent, then X and Y both have normal distributions; Cramèr's theorem), and also for the Poisson distribution (same result; Raikov's theorem).

Finally, if we turn to the question of factorizations of infinitely divisible distributions *which remain in the ambit of infinitely divisible distributions* (i.e. we require that the factors also be such), we can say straightaway that in this case the answer is straightforward and complete. We shall deal with this in Chapter 8, 8.4 (at the present time we do not have at our disposal the concepts required for taking this any further).

It is instructive to point out the following rather surprising fact: given a factorization $\phi(u) = \phi_1(u)\phi_2(u)$, this does not imply that if one factor is kept fixed the other is uniquely determined (in other words, we can also have $\phi(u) = \phi_1(u)\phi_3(u)$, with $\phi_3 \neq \phi_2$). Clearly, we can only have $\phi_3(u) \neq \phi_2(u)$ when $\phi_1(u) = 0$; but we have already seen that a characteristic function can be zero (like the triangular case, $1 - \frac{1}{2}|u|$; see 6.11.3) outside an interval (in this example, for $|u| \geqslant 2$). In fact, the counter example given by Khintchin consists precisely in taking ϕ_1 to be such a triangular function; for ϕ_2 and ϕ_3, one can take, for example, concave polygonal functions (see Pólya's theorem) which are the same in $(-2 \leqslant u \leqslant 2)$ but differ outside.