7

A Preliminary Survey

7.1 Why a Survey at this Stage?

7.1.1. Our discussion of the requirements of the conceptual formulation of the theory of probability has already revealed its wide range of application. It applies, in fact, whenever the factor of uncertainty is present. The range of problems encountered is also extensive. Diverse in nature and in complexity, these problems require a corresponding range of mathematical techniques for their formulation and analysis, techniques which are provided by the calculus of probability. For a number of reasons, it is useful to give a preliminary survey, illustrating these various aspects. In setting out our reasons, and by inviting the reader to take note of certain things, we shall be able to draw attention to those points which merit and require the greatest emphasis.

First of all, we note that individual topics acquire their true status and meaning only in relation to the subject as a whole. This is probably true of every subject, but it is particularly important in the case of probability theory. In order to explore a particular area, it pays to get to know it in outline before starting to cover it in great detail (although this will be necessary eventually), so that the information from the detailed study can be slotted into its rightful place. If we were to proceed in a linear fashion, we would not only give an incomplete treatment but also a misleading one, in that it would be difficult to see the connections between the various aspects of the subject. The same would be true of even the most straightforward problems if we had to deal with them without, at any given point, referring to any feature whose systematic treatment only came later (e.g. we would not be able to mention the connections with 'laws of large numbers', 'random processes' or 'inductive inference'). Nor would it be reasonable (either in general or in this particular case) to assume that individual chapters are approached only after all the preceding ones have been read, and their contents committed to memory. For an initial appreciation, it is necessary and sufficient to be clear about basic problems and notions rather than attempting to acquire a detailed knowledge. Moreover, the difficulties associated with this approach are easily avoided. It is sufficient to learn from the outset how to *understand* what these problems and concepts are about by concentrating on a small number of simple but meaningful examples. Although elementary and summary in nature, the approach is then both clear and concrete, and can be further developed by various additional comments and information.

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

7.1.2. In this preliminary survey, we shall, for this reason, concentrate on the case of *Heads and Tails*. This example, examined from all possible angles, will serve as a basic model, although other variants will be introduced from time to time (more for the sake of comparison and variety than from necessity). These simple examples will shed light on certain important ideas that crop up over and over again in a great many problems, even complex and advanced ones. This, in turn, facilitates the task of analysing the latter in greater depth. In fact, it often turns out that the result of such an analysis is simply the extension of known and intuitive results to those more complex cases. This also reveals that the detailed complications which distinguish such cases from the simple ones are essentially irrelevant.

Another reason for providing a survey is the following. Everybody finds problems in the calculus of probability difficult (nonmathematicians, mathematicians who are unfamiliar with the subject, even those who specialize in it if they are not careful[1]). The main difficulty stems, perhaps, from the danger of opting for the apparently obvious, but wrong, conclusion, whereas the correct conclusion is usually easily established, provided one looks at the problem in the right way (which is not – until one spots it – the most obvious). In this respect, the elementary examples provide a good basis for discussion and advice (which, although useful, is inadequate unless one learns how to proceed by oneself for each new case). Many of the comments we shall make, however, are not intended solely for the purpose of avoiding erroneous or cumbersome arguments when dealing with simple cases. More generally, they are made with the intention of clarifying the conceptual aspects themselves, and of underlining their importance, in order to avoid any misunderstandings or ambiguities arising in other contexts going beyond those of the examples actually used. In other words, we shall be dealing with matters which, as far as the present author is concerned, have to be treated as an integral part of the formulation of the foundations of the subject, and which could have been systematically treated as such were it not for the risk that one might lose sight of the direct nature of the actual results and impart to the whole enterprise a suggestion of argument for argument's sake, or of literary–philosophical speculation.

7.1.3. It turns out that the aims that we have outlined above are best achieved by concentrating mainly on examples of the 'classical' type – that is those based on combinatorial considerations. In fact, even leaving aside the need to mention such problems anyway, many of these combinatorial problems and results are particularly instructive and intuitive by virtue of their interpretations in the context of problems in probability. Indeed, it was once thought that the entire calculus of probability could be reduced essentially to combinatorial considerations by reducing everything down to the level of 'equiprobable cases'. Although this idea has now been abandoned, it remains true that combinatorial

[1] Feller, for example, repeatedly remarks on the way in which certain results seem to be surprising and even paradoxical (even simple results concerning coin tossing, such as those relating to the periods during which one gambler has an advantage over the other; Chapter 8, 8.7.6). He remarks on 'conclusions that play havoc with our intuition' (p. 68), and that 'few people will believe that a perfect coin will produce preposterous sequences in which no change of lead occurs for millions of trials in succession, and yet this is what a good coin will do rather regularly' (p. 81). Moreover, he can attest to the fact that 'sampling of expert opinion has revealed that even trained statisticians feel that' certain data are really surprising (p. 85).

All this is from W. Feller, *An Introduction to Probability Theory and its Applications*, 2nd edn, John Wiley & Sons, Inc., New York (1957). Many of the topics mentioned in the present chapter are discussed in detail by Feller (in Chapter 3, in particular), who includes a number of original contributions of his own.

considerations do play an important rôle, even in cases where such considerations are not directly involved (see the examples that were discussed in Chapter 5, 5.7.4).

Given our stated purpose in this 'preliminary survey', it will naturally consist more of descriptive comments and explanations than of mathematical formulations and proofs (although in cases where the latter are appropriate we shall provide them). In the first place, we shall deal with those basic, straightforward schemes and analyses which provide the best means of obtaining the required 'insights'. Secondly, we shall take the opportunity of introducing (albeit in the simplest possible form) ideas and results that will be required in later chapters, and of subjecting them to preliminary scrutiny (although without providing a systematic treatment). Finally, we shall consider certain rather special results which will be used later (here they link up rather naturally with one of the examples, whereas introduced later they would appear as a tiresome digression).

7.2 Heads and Tails: Preliminary Considerations

7.2.1. Unless we specifically state otherwise, we shall, from now on, be considering events which You judge to have probability $\frac{1}{2}$ and to be stochastically independent. It follows that each of the 2^n possible results for n such events all have the same probability, $(\frac{1}{2})^n$ Conversely, to judge these 2^n results to be all equally probable implies that You are attributing probability $\frac{1}{2}$ to each event and judging the events to be stochastically independent.

The events $E_1, E_2, \ldots, E_m, \ldots$ will consist of obtaining Heads on a given toss of a coin (we could think in terms of some preassigned number of tosses, n, or of a random number – for example 'until some specified outcome is obtained', 'those tosses which are made today' and so on – or of a potentially infinite number). We shall usually take it that we are dealing with successive tosses of the same coin (in the order, $E_1, E_2, \ldots$), but nothing is altered if one thinks of the coin being changed every now and then, or even after every toss. In the latter case, we could be dealing with the simultaneous tossing of n coins, rather than n successive tosses (providing we establish some criterion other than the chronological one – which no longer exists – for indexing the E_i). We could, in fact, consider situations other than that of coin tossing. For example: obtaining an even number on a roll of a die, or at bingo; or drawing a red card from a full pack, or a red number at roulette (excluding the zero) and so on. We shall soon encounter further examples, and others will be considered later.

7.2.2. In order to represent the outcomes of n tosses (i.e. a sequence of n outcomes resulting in either Heads or Tails), we can either write *HHTHTTTHT*, or, alternatively, 110100010 (where Heads = 1, Tails = 0).[2]

A. How many Heads appear in the n tosses? This is the most common question. We know already that out of the 2^n possibilities the number in which Heads appear h times is given by $\binom{n}{h}$. The probability, $\omega_h^{(n)}$ of h successes out of n events is therefore $\binom{n}{h}/2^n$. We shall return to this question later, and develop it further.

[2] It should be clear that expressions like *HHT* (denoting that three consecutive outcomes – e.g. the first, second and third, or those labelled n, $n + 1$, $n + 2$ – are Head–Head–Tail) are merely suggestive 'shorthand' representations. The actual logical notation would be $E_n E_{n+1} \tilde{E}_{n+2}$ (or $H_n H_{n+1}\ T_{n+2}$, if one sets $H_i = E_i$ and $T_i = \tilde{E}_i$). Let everyone be clear about this, so that no one inadvertently performs operations on *HHT* as though it were simply a product (it would be as though one thought that the year 1967, like *abed*, being the product of four 'factors', that is 1, 9, 6, 7, were equal to 378).

B. How many runs of consecutive, identical outcomes are there? In the sequence given above, there were six runs: *HH*/*T*/*H*/*TTT*/*H*/*T*. It is clear that after the initial run we obtain a new run each time an outcome differs from the preceding one. The probability of obtaining $h + 1$ runs is, therefore, simply that of obtaining *h change-overs*, and so we consider:

C. How many change-overs are there? In other words, how many times do we obtain an outcome which differs from the preceding one? For each toss, excluding the initial one, asking whether or not the toss gives the same outcome as the previous one is precisely the same as asking whether it gives Heads or Tails. The question reduces, therefore, to (*A*), and the probability that there are *h* change-overs in the *n* tosses is equal to $\binom{n-1}{h}/2^{n-1}$.

D. Suppose we know that out of $n = r + s$ tosses, *r* are to be made by Peter and *s* by Paul. *What is the probability that they obtain the same number of successes*? Arguing systematically, we note that the probability of Peter obtaining *h* successes and Paul obtaining *k* is equal to $\binom{r}{h}\binom{s}{k}/2^n$. It follows that the probability of each obtaining the same number of successes is given by $\left(\frac{1}{2}\right)^n \sum_h \binom{r}{h}\binom{s}{h}$ (the sum running from 0 to the minimum of *r* and *s*). As is well-known, however (and can be verified directly by equating coefficients in $(1 + x)^r. (1 + x)^s = (1 + x)^{r+s}$), this sum is equal to $\binom{n}{r} = \binom{n}{s}$, and the probability that we are looking for is identical to that of obtaining *r* (or *s*) successes out of *n* tosses.

This result could have been obtained in an intuitive manner, and without calculation, by means of a similar device to that adopted in the previous case. We simply note that the problem is unchanged if 'success' for Paul is redefined to be the outcome Tails rather than Heads. To obtain the same number of successes (*h* say) now reduces to obtaining *s* Heads and *r* Tails overall; $s = h + (\mathrm{s} - h)$, $r = (r - h) + h$. Without any question, this is the most direct, natural and instructive *proof* of the combinatorial identity given above.

E. What is the probability that the number of successes is odd? There would be no difficulty in showing this to be $\frac{1}{2}$, by plodding through the summation of the binomial coefficients involved (the sums of those corresponding to evens and odds are equal!). If *n* were odd, it would be sufficient to observe that an odd number of Heads entails an even number of Tails, and so on.

A more direct and intuitive argument follows from noting that we need only concern ourselves with the final toss. The probability of a success is $\frac{1}{2}$ (no matter what happened on the preceding tosses), and hence the required probability is $\frac{1}{2}$. The advantage of this argument is that we see, with no further effort, that the same conclusion holds under much weaker conditions. It holds, in fact, for any events whatsoever, logically or stochastically independent, and with arbitrary probabilities, provided that one of them has probability $\frac{1}{2}$, and is independent of all combinations of the others (or, at least, of the fact of whether an odd or an even number of them occur).[3]

We shall return to this topic again in Section 7.6.9.

[3] Pairwise independence (which we consider here in order to show how much weaker a restriction it is) would not entitle us to draw these conclusions. We can obtain a counterexample by taking just three events, *A*, *B*, *C*, and supposing them all to be possible, with the four events 'only *A*' 'only *B*' 'only *C*' and 'all three' (*ABC*) equally probable ($p = \frac{1}{4}$). It is easily seen that *A*, *B*, *C* each have probability $\frac{1}{2}$ and are pairwise independent, but that the number of successes is certainly odd (either 1 or 3). If we had argued in terms of the complements, it would certainly be even (either 0 or 2).

7.2.3. *Some comments.* The main lesson to be learned from these examples is the following. *In the calculus of probability, just as in mathematics in general, to be able to recognize the essential identity of apparently different problems is not only of great practical value but also of profound conceptual importance.*

In particular, arguments of this kind often enable us to avoid long and tedious combinatorial calculations; indeed, *they constitute the most intuitive and 'natural' approach to establishing combinatorial identities.*[4] Moreover, they should serve, from the very beginning, to dispel any idea that there might be some truth in *certain of the specious arguments one so often hears repeated.* For example: that there is some special reason (in general, that it is advantageous) to either always bet on Heads, or always on Tails; or, so far as the lottery is concerned, to always bet on the same number, perhaps one which has not come up for several weeks! All this, despite the fact that, by assumption, all the sequences are equally probable. It is certainly true that the probability of no Heads in ten successive tosses is about one in a thousand ($2^{-10} = 1/1024$), and in twenty tosses about one in a million ($2^{-20} = 1/1048576$), but the fact of the matter is that the probability of not winning in ten (or twenty) tosses if one always sticks to either Heads or Tails is always exactly the same (that given above). This is the case no matter whether or not the tosses are consecutive, or whether or not one always bets on the same face of the coin, or whether one alternates in a regular fashion, or decides randomly at every toss. To insist on sticking to one side of the coin, or to take the consecutive nature of the tosses into account, is totally irrelevant.

7.2.4. *F. What is the probability that the first* (*or, in general, the* rth) *success* (*or failure*) *occurs on the* hth *toss*? The probability of the first success occurring on the hth toss is clearly given by $(\frac{1}{2})^h$ (the only favourable outcome out of all the 2^h is given by 000...0001). Note that this probability, $(\frac{1}{2})^h$, is the same as that of obtaining *no successes in h tosses*; that is of having to perform more than h tosses before obtaining the first success.[5] The probability of the rth success occurring at the hth toss is given by $(\frac{1}{2})^h\binom{h-1}{r-1}$, because this is the probability of exactly $r-1$ successes in the first $h-1$ tosses multiplied by the probability $(\frac{1}{2})$ of a further success on the final (hth) toss.

G. A coin is alternatively tossed, first by Peter and then by Paul, and so on. *If the winner is the one who first obtains a Head, what are their respective probabilities of winning*? A dull, long-winded approach would be to sum the probabilities $(\frac{1}{2})^h$ for h odd (to obtain Peter's probability of winning), or h even (for Paul's probability), and this would present no difficulties. The following argument is more direct (although its real advantage shows up better in less trivial examples). If Peter has probability p, Paul must have

[4] My 'philosophy' in this respect is to consider as a *natural proof* that which is based on a combinatorial argument, and as a more or less *dull verification* that which involves algebraic manipulation. In other fields, too, certain things strike me as mere 'verifications'. For example: proofs of vectorial results which are based upon components; properties of determinants established by means of expansions (rather than using the ideas of alternating products, volume or, as in Bourbaki, smoothly generated by means of an exterior power). Indeed, this applies to anything which can be proved in a synthetic, direct and (meaningfully) instructive manner, but which is proved instead by means of formal machinery (useful for the bulk of the theory, but not for sorting out the basic ideas).

[5] We observe that the probability of no successes in n tosses tends to zero as n increases. This is obvious, but it is necessary to draw attention to it, and to make use of it, if certain arguments are to be carried through correctly (see the *Comments* following (G)).

probability $\tilde{p} = \frac{1}{2}p$, because he will find himself in Peter's shoes if the latter fails to win on the first toss; we therefore have $p = \frac{2}{3}, \tilde{p} = \frac{1}{3}$.

Comments. We have tacitly assumed that one or other of them certainly ends up by winning. In actual fact, we should have stated beforehand that, as we pointed out in the footnote to (*F*), the probability of the game not ending within n tosses tends to zero as n increases. In examples where this is not the case, the argument would be wrong.

7.2.5. *H. What is the probability of obtaining, in n tosses, at least one run of h successes* (i.e. at least h consecutive successes)? Let A_n denote the number of possible sequences of n outcomes which do not contain any run of h Heads. By considering one further trial, one obtains a set of $2A_n$ sequences, which contains all the A_{n+1} sequences with no run of h Heads in $n + 1$ trials, plus those sequences where the last outcome – which is therefore necessarily a Head – forms the first such sequence. There are A_{n-h} of these, because they must be obtained by taking any sequence of $n - h$ trials with no run of h Heads, and then following on with a Tail and then h Heads. We therefore obtain the recurrence relation $A_{n+1} = 2A_n - A_{n-h}$, in addition to which we know that $A_0 = 1, A_n = 2^n$ for $n < h$, $A_h = 2^h - 1$ and so on.

We are dealing here with a difference equation. It is well known (and easy to see) that it is satisfied by x^n, where x is a root of the (characteristic) equation $x^{h+1} - 2x^h + 1 = 0$. The general solution is given by

$$A_n = a_0 + a_1 x_1^n + a_2 x_2^n + \ldots + a_h x_h^n,$$

where 1, $x_1, \ldots, x_h$ are the $h + 1$ roots, and the constants are determined by the initial conditions.

We shall confine attention to the case $h = 2$. The recurrence relation $A_{n+1} = 2A_n - A_{n-2}$ can be simplified[6] so that it reduces to that of the Fibonacci numbers (each of which is the sum of the two preceding ones); that is $A_{n+1} = A_n + A_{n-1}$. In fact, however, a direct approach is both simpler and more meaningful. Those of the A_{n+1} sequences ending in Tails are the A_n followed by a Tail; those ending in Heads are the A_{n-1} followed by Tail–Head; the formula then follows immediately.

Using the fact that $A_0 = 1, A_1 = 2, A_2 = 3$, we find that $A_3 = 5, A_4 = 8, A_5 = 13, A_6 = 21, A_7 = 34, A_8 = 55$, and so on, and hence that the required probability is $1 - A_n/2^n$. For four trials this gives $1 - \frac{8}{16} = \frac{1}{2}$, for eight trials $1 - \frac{55}{256} = 0 \cdot 785$, and so on. To obtain the analytic expression, we find the roots of $x^2 - x - 1 = 0$ (noting that

$$x^3 - 2x^2 + 1 = (x-1)(x^2 - x - 1) = 0),$$

obtaining $x_{1,2} = (1 \pm \surd 5)/2$, and hence

$$A_n = \left[\left(1+\surd 5\right)^{n+1} - \left(1-\surd 5\right)^{n+1}\right] \Big/ 2^{n+1}\surd 5.$$

A similar argument will work for any $h > 1$. The A_{n+1} are of the form $A_n T$, A_{n-1} *TH*, A_{n-2} *THH*,..., A_{n-h+1} *THHH ... H* (with $h - 1$ Heads), where the notation conveys that an A_{n-k} is followed by a Tail and then by k Heads. It follows that A_{n+1} = the sum of the h preceding terms (and this is clearly a kind of generalization of the Fibonacci condition).

[6] By writing it in the form $A_{n+1} - A_n - A_{n-1} = A_n - A_{n-1} - A_{n-2}$, we see that the expression is independent of n; for $n = 2$, we have $A_2 - A_1 - A_0 = 3 - 2 - 1 = 0$.

The equation one arrives at ($1 + x + x^2 + x^3 + \ldots + x^h = x^{h+1}$) is the same as the one above, divided by $x - 1$ (i.e. without the root $x = 1$).

Comments. We have proceeded by *induction*; that is with a *recursive* method. This is a technique that is often useful in probability problems – keep it in mind!

Note that in considering the A_{n-h} we have discovered in passing the probability that a run of h successes is completed for the first time at the (n + l)th trial. This is given by $A_{n-h}/2^{n+1}$. It is always useful to examine the results that become available as byproducts. Even if they do not seem to be of any immediate interest, they may throw light on novel features of the problem, suggest other problems and subsequently prove valuable (You never know!).

On the other hand, this probability (i.e. that something or other occurs for the first time on the (n + 1)th trial) can always be obtained by subtracting the probability that it occurs at least once in the first $n + 1$ trials from the probability that it occurs at least once in the first n (or by subtracting the complementary probabilities).

This remark, too, is obvious, but it is important, nonetheless. It often happens that the idea is not used, either because it is not obviously applicable, or because it simply does not occur to one to use it.

7.2.6. *I. What is the probability that a particular trial (the* nth *say) is preceded by exactly h outcomes identical to it, and followed by exactly k*? In other words, what is the probability that it forms part of a run of (exactly) $h + k + 1$ identical outcomes (either all Heads or all Tails) of which it is the (h + 1)th (we assume that $h < n - 1$).[7] The probability is, in fact, equal to $(\frac{1}{2})^{h+k+2}$. We simply require the h previous outcomes and the k following to be identical, and the outcome of the trial preceding this run, and the one following it, to be different (in order to enclose the run).

J. What is the probability that the nth *trial forms part of a run of (exactly) m trials having identical outcomes*? This, of course, reduces to the previous problem with h and k chosen such that $h + k + 1 = m$ (naturally, we assume $m \leqslant n - 1$). For each individual possible position, the probability is $(\frac{1}{2})^{m+1}$, and there are m such cases since the nth trial could either occupy the 1st, 2nd,..., or the mth position in the run (i.e. we must have one of

$$h = 0,\ 1,\ 2,\ \ldots,\ m-1).$$

The required probability is therefore $m/2^{m+1}$.

In particular, the probability of a particular trial being *isolated* (i.e. a Tail sandwiched between two Heads, or vice versa) is equal to $\frac{1}{4}$; the same is true for a run of length two, and we have $\frac{3}{16}$ for a run of length three, $\frac{4}{32} = \frac{1}{8}$ for a run of length four, $\frac{5}{64}$ for a run of length five and so on.

K. What is the probability that some 'given' run, the nth *say, has (exactly) length m*? The nth run commences with the (n – l)th change-over, and has length m if the following $m - 1$ outcomes are identical and the mth is different. The probability of this is given by $(\frac{1}{2})^m$. For lengths 1, 2, 3, 4, 5, we therefore have probabilities $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}$, and so on.

[7] We exclude the (possible) case $h = n - 1$, which would give a different answer ($(\frac{1}{2})^{h+k+1}$: Why?).

Comments. It might appear that (*J*) and (*K*) are asking the same question: that is in both cases one requires to know the probability that a run (or 'a run *chosen at random*') has length m. The problem is not well defined, however, until *a particular* run has been specified. The two methods of doing so – on the one hand demanding that the run contain some given element, on the other hand that it be the run with some given label – lead to different results, yet both methods could claim to be 'choosing a run at random'. We shall often encounter 'paradoxes' of this kind and this example (together with the developments given under (*L*)) serves precisely to draw attention to such possibilities.

Warning. All the *relevant* circumstances (*and, at first sight, many of them often do not seem relevant*!) must be *set out very clearly* indeed, in order to avoid essentially different problems becoming confused. The phrase 'chosen at random' (and any similar expression) *does not, as it stands, have any precise meaning*. On the contrary, assuming it to have some uniquely determined intrinsic meaning (which it does not possess) is a common source of error. Its use is acceptable, however, provided it is always understood as indicating something which subsequently has to be made precise in any particular case. (For example, it may be that at some given instant a person decides that 'choosing a run at random' will have the meaning implicit in (J), or it may be that he decides on that of (*K*), or neither, preferring instead some other interpretation.) In order not to get led astray by the overfamiliar form of words, one might substitute in its place the more neutral and accurate form 'chosen in some quite natural and systematic way (which will be made precise later)'.[8]

7.2.7. *L*. *What is the prevision of the length of a run* (*under the conditions given in* (*J*) *and* (*K*), *respectively*)? Let us begin with (*K*). It might appear that the random quantity L = 'length of the run' can take on possible values 1, 2,..., m,... with probabilities $\frac{1}{2}$, $\frac{1}{4}$,..., $(\frac{1}{2})^m$,... and that, therefore, its prevision is given by $\sum m/2^m$. But are we permitting ourselves the use of the series, considering the sequence as infinitely long (a possibility we previously excluded, for the time being anyway), or should we take the boundedness of the sequence into account (by assuming, for instance, that the number of trials does not exceed some given *N*)? It seems to be a choice between the devil and the deep blue sea, but we can get over this by thinking of *N* as finite, but large enough for us to ignore the effect of the boundedness. In other words, we accept the series in an unobjectionable sense; that is as an asymptotic value as *N* increases. Although in this particular case the series $\sum mx^m = x/(1-x)^2$ presents no difficulties (we see immediately that it gives $\mathbf{P}(L) = 2$), there are often more useful ways of proceeding. Given that prevision is additive, and given that the length L is the sum of as many 1s as there are consecutive identical outcomes (the first of which is certain, the others having probabilities $\frac{1}{2}$,$\frac{1}{4}$, ..., etc., as we saw above), we obtain

$$\mathbf{P}(L) = 1 + \frac{1}{2} + \frac{1}{4} + \ldots + \left(\frac{1}{2}\right)^m + \ldots = 2.$$

[8] See, for instance, the remarks in Chapter 5, 5.10.2 concerning the notion of 'equiprobable' in quantum physics, and those in Chapter 10, 10.4.5 concerning the 'random choice' of subdivision point of an interval, or in a Poisson process.

Alternatively, and even more directly (using an argument like that in (*G*)), we note that we must have $\mathbf{P}(L) = 1 + \frac{1}{2}\mathbf{P}(L)$ (and hence $\mathbf{P}(L) = 2$), since, if the second outcome is the same as the first (with probability $\frac{1}{2}$), the length of the run starting with it has the same prevision $\mathbf{P}(L)$.

Going through the same process in case (*J*), we would have

$$\mathbf{P}(L') = \sum m(m/2^{m+1}) = \sum m^2/2^{m+1} = x(1+x)/2(1-x)^3 = 3\left(\text{for } x = \frac{1}{2}\right),$$

but the argument could be simplified even more by recalling that L had the value 1 + the prevision of the identical outcomes to the right (which was also 1), and that here we should add the same value 1 as a similar prevision to the left, so that we obtain 1 + 1 + 1 = 3. Of course, we must also assume *n* large, but, in any case, it would be easy to evaluate the rest of the series in order to put bounds on the error of the asymptotic formula were the accuracy to be of interest.

It turns out, therefore, that the previsions resulting from the two different methods of choosing the runs are different, 2 and 3, respectively (as one might have expected, given that the smaller values were more probable in the first case, and conversely for the others). If we ignore the certain value 1 for the initial outcome, the additional length turns out to be double in the second case (2 instead of 1), because the situation is the same on both sides (independently of the fact that there is no actual continuation to the left, since, by hypothesis, the initial term is the first one in the run; this in no way changes the situation on the right, however: '*the later outcomes neither know nor care about this fact*').

Comment. A sentence like the above, or, equivalently, 'the process has no memory' (as in the case of stochastic independence), is often all that is required to resolve a paradox, or to avoid mistakes (like those implicit in the *well-known specious arguments* which we mentioned in our Comments following (*E*)).

M. Suppose we toss a coin n times: what are the previsions of the number of successes (Heads);
change-overs (Head followed by Tail, or vice versa);
runs;
runs of length m;
*tosses up to and including the h*th *success;*
tosses up to and including the completion for the first time of a run of successes of length 2 (*or, in general, of length h*)?

Most questions of this kind are much easier than the corresponding questions involving probabilities (as one would expect, given the additivity of prevision).

So far as successes are concerned, at each toss the probability of success is $\frac{1}{2}$, and hence the prevision of the number of successes in *n* tosses is $\frac{1}{2}n$.

For the change-overs, the same argument applies (apart from the case of the first toss), and we have $\frac{1}{2}(n-1)$.

For runs, we always have 1 more than the number of change-overs, and hence the prevision is $\frac{1}{2}(n+1)$.

For runs of length *m*, we shall give, for simplicity, the asymptotic expression for *n* large in comparison to *m* (see (*L*)). For each toss (and, to be rigorous, we should modify this for the initial and final m tosses), we have probability $m/2^{m+1}$ of belonging to a run of length m, and so, in prevision, there are $nm/2^{m+1}$ such tosses out of n; there are, therefore, $n/2^{m+1}$ runs of this kind (since each consists of m tosses). In particular, we have, in prevision, $n/4$ isolated outcomes, $n/8$ runs of length two, and so on.

From (F), we can say that the prevision of the number of tosses required up to and including the hth success is given by $\sum k\binom{k-1}{h-1}/2^k$ (the sum being taken over 1 to ∞, and, as usual, thought of as an asymptotic value). It is sufficient, however, to restrict attention to the very simple case $h = 1$. The prevision of the number of tosses required for the first success is 2 (the probability of it occurring on the first toss is $\frac{1}{2}$, at the second $\frac{1}{4}$, etc.); that is it is given by $\sum m/2^m$ as in the first variant discussed under (L). It follows that the prevision of the number of tosses required for the hth success is $2h$ (and that this is therefore the value of the summation given above, a result which could be verified directly).

For the final problem, we note that the probability of the first run of two Heads being completed at the nth toss is given by $A_{n-2-1}/2^n$ (where the A denote Fibonacci numbers), and that the prevision is therefore given by $\sum nA_{n-3}/2^n$. There is a useful alternative method of approach, however. Let us denote this prevision by $\mathbf{P}(L)$, and note the following: if the first two tosses both yield Heads, we have $L = 2$ (and this is the end of the matter); if they yield Head–Tail, we have $\mathbf{P}(L|HT) = 2 + \mathbf{P}(L)$, because the situation after the first two tosses reverts back to what it was at the beginning; if the first toss results in a Tail, we have, similarly, $\mathbf{P}(L|T) = 1 + \mathbf{P}(L)$. Since the probabilities of these three cases are $\frac{1}{4}, \frac{1}{4}, \frac{1}{2}$, respectively, we obtain

$$\mathbf{P}(L)=\left(\frac{1}{4}\right).2+\left(\frac{1}{4}\right)\left(\mathbf{P}(L)+2\right)+\left(\frac{1}{2}\right)\left(\mathbf{P}(L)+1\right)=\frac{3}{2}+\frac{3}{4}\mathbf{P}(L),$$

which implies that $\left(\frac{1}{4}\right)\mathbf{P}(L)=\frac{3}{2}$, and hence that $\mathbf{P}(L) = 6$.

The argument for the first run of h Heads proceeds similarly. Let us briefly indicate how it goes for the case $h = 3$: first three tosses Head–Head–Head, probability $\frac{1}{8}$, $L = 3$; first three tosses Head–Head–Tail, probability $\frac{1}{8}$, $\mathbf{P}(L|HHT) = 3 + \mathbf{P}(L)$; first two tosses Head–Tail, probability $\frac{1}{4}$, $\mathbf{P}(L|HT) = 2 + \mathbf{P}(L)$; first toss Tail, probability $\frac{1}{2}$, $\mathbf{P}(L|T) = 1 + \mathbf{P}(L)$. Putting these together, we obtain $\mathbf{P}(L) = 14$. For $h = 4$, we obtain $\mathbf{P}(L) = 30$; the general result is given by $\mathbf{P}(L) = 2^{h+1} - 2$ (prove it!).

7.2.8. *Remarks*. Let us quickly run through some other possible interpretations, and, in so doing, draw attention to certain features of interest. Instead of simply dealing with the Head and Tail outcomes themselves, we could consider their 'matchings' with some given 'comparison sequence', $E_1^*, E_2^*, \ldots, E_n^*, \ldots$. For example, if the comparison sequence were chosen to be the alternating sequence *HTHTHT…*, and we used 1 to denote a matching, 0 otherwise, then we obtain a 1 whenever a Head appears on an odd toss, or a Tail on an even. In this way, any problem concerning 'runs' can be reinterpreted directly as one concerning 'alternating runs'. The comparison sequence could be the sequence in which a gambler 'bets' on the outcome of the tosses: for example, *HHTHTTTHT* = 'he bets on Heads at the 1st, 2nd, 4th and 8th tosses, and he bets on Tails at the 3rd, 5th, 6th, 7th and 9th tosses'. The outcomes 1 and 0 then denote that 'he wins' or 'he loses', respectively. It follows, in this case, with no distinction drawn between Heads or Tails, that 'runs' correspond to runs of wins or losses, whereas if 'runs' refer to

Heads only, say (as in (*H*)), then they correspond to runs of wins only (and conversely, if 'runs' refer to Tails only). The comparison sequence could even be random. For example, it might have arisen as the result of another sequence of coin tosses, or from some other game or experiment (double six when rolling two dice; the room temperature being below 20°C; whether or not the radio is broadcasting music; whether at least $\frac{1}{3}$ of those present have blond hair, etc.). This other experiment may or may not be performed simultaneously and could depend on the outcomes of the sequence itself (for example, if $E_n^* = \tilde{E}_{n-1}$ we obtain the case of 'change-overs' as in (*C*)). The only condition is that E_h^* be stochastically independent of E_h (for each h). In fact, if, conditional on any outcome for the $E_i(i \neq h)$, E_h has probability $\frac{1}{2}$, then the same holds for $(E_h^* = E_h)$, no matter what the event E_h^* is, provided that it is independent of E_h (this can be seen as a special case of (*E*) for $n = 2$, but there it is obvious anyway).

7.3 Heads and Tails: The Random Process

7.3.1. In Section 7.2, we confined ourselves to a few simple problems concerning the calculation of certain probabilities and previsions in the context of coin tossing. This provided a convenient starting point for our discussion, but now we wish to return to the topic in a more systematic manner. In doing so, we shall get to know many of the basic facts, or, at least, become acquainted with some of them, and we shall also encounter concepts and techniques that will later come to play a vital rôle. In particular, we shall see how second-order previsions often provide a fruitful way of getting at important results and we shall encounter various distributions, random[9] processes, asymptotic properties and so on. Let us proceed straightaway to a consideration of why it is useful, even when not strictly necessary, to formulate and place these problems in a *dynamic* framework, as *random processes*, or as *random walks*.

An arbitrary sequence of events $E_1, E_2,\dots, E_n,\dots$ (which, unless we state otherwise, could be continued indefinitely) can, should one wish to do so, be considered as already constituting in itself a random function, $E_n = Y(n)$,[10] assigning either 1 or 0 to each positive integer n. To obtain more meaningful representations, one could, for instance, consider the *number of successes* $Y(n) = S_n = E_1 + E_2 + \dots + E_n$, or the *excess of successes* over failures

$$Y(n) = 2S_n - n = \left(E_1 + E_2 + \dots + E_n\right) - \left(\tilde{E}_1 + \tilde{E}_2 + \dots + \tilde{E}_n\right).$$

The latter could also be considered as the *total gain*,

$$Y(n) = X_1 + X_2 + \dots + X_n,$$

if $X_i = 2E_i - 1 = E_i - \tilde{E}_i$ is defined to be the gain at each event; that is $X_i = \pm 1$ (one gains 1 or loses 1 depending on whether E_i is true or false; put in a different way, one always pays 1 and receives 2 if the event occurs, receives 1 and pays 2 if it does not).

[9] See Chapter 1, 1.10.2, for a discussion of the use of the words 'random' and 'stochastic'.

[10] We shall usually write $Y(n)$, $Y(t)$, only when the variable (e.g. time) is *continuous*. When it is *discrete* we shall simply write Y_n, Y_t, except, as here, when we wish to emphasize that we are thinking in terms of the *random process*, rather than of an *individual* Y_n.

If it should happen (or if we make the assumption) that events occur after equal (and unit) time intervals, then we can say that $Y(t)$ is the number (or we could take the excess) of successes up to time t (i.e. $Y(t) = Y(n)$ if $t = n$, and also if $n \leqslant t < n + 1$). For the time being, this merely serves to provide a more vivid way of expressing things (in terms of 'time' rather than 'number of events'), but later on it will provide a useful way of showing how one passes from processes in *discrete time* to those in *continuous time* (even here this step would not be without meaning if events occurred at arbitrary time instants $t_1 < t_2 < \ldots < t_n < \ldots$, especially if randomly, as, for instance, in the Poisson process; Chapter 8, 8.1.3).

7.3.2. The representation which turns out to be most useful, and which, in fact, we shall normally adopt, is that based on the *excess* of successes over failures, $Y(n) = 2S_n - n$. This is particularly true in the case of coin tossing, but to some extent holds true more generally.

The possible points for $(n, Y(n))$ in the (t, y)-plane are those of the lattice shown in Figure 7.1a. They have integer coordinates, which are either both even or both odd ($t = n \geqslant 0, -t \leqslant y \leqslant t$). It is often necessary, however, to pick out the point which corresponds to the number of successes in the first n events, and, in order to avoid the notational inconvenience of $(n, 2h - n)$, we shall, by convention, denote it by $[n, h]$: in other words, $[t, z] = (t, 2z - t)$ $(t, y) = [t, \frac{1}{2}(y + t)]$. As can be seen from Figure 7.1b, this entails referring to the coordinate system (t, z), with vertical lines (t = constant), and downward sloping

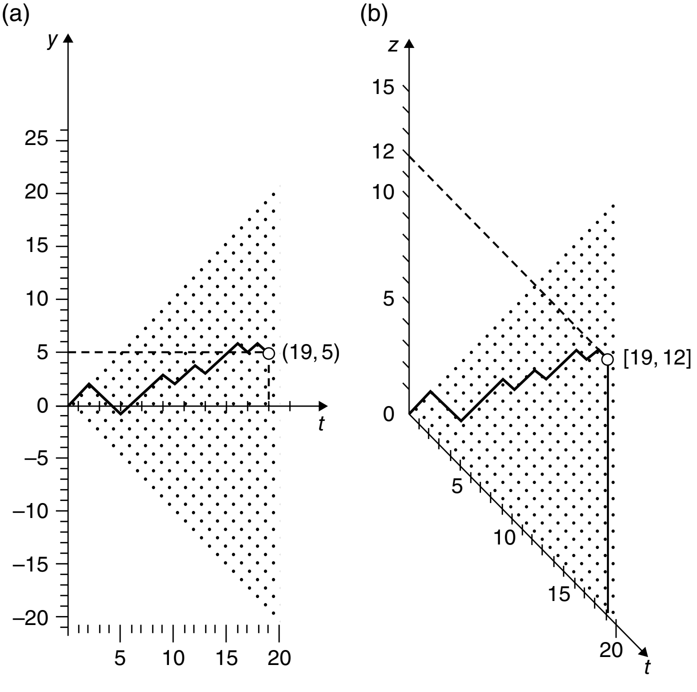

Figure 7.1 The lattice of possible points for the coin-tossing process. Coordinates: time $t = n$ = number of tosses (in both cases), together with: in (a): y = gain = $h - (n - h)$ = number of successes minus number of failures; in (b): $z = h$ = number of successes = $(n + y)/2$. The notations (t, y) and $[t, z]$ refer, respectively, to the coordinate systems of (a) and (b). For the final point of the path given in the diagrams, we have, for example,

$$(19,5) \equiv [19,12] \quad (t = n = 19, h = 12, n - h = 7, y = 12 - 7 = 5)$$

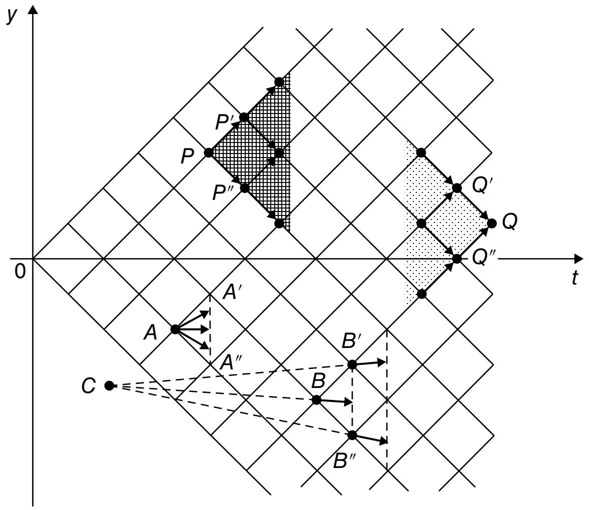

Figure 7.2 The lattice of the 'random walk' for coin tossing (and similar examples). The point *Q* can only be reached from *Q*′ or *Q*″ (this trivial observation often provides the key to the formulation of problems). It follows, therefore, that it can only be reached from within the angular region shown. Similarly, the other angular region shows the points that can be reached from *P*. The vectorial representation at A provides a way of indicating the probabilities of going to *A*′ rather than to *A*″ (reading from the bottom upwards, we have probabilities $p = 0{\cdot}2$, 0·5, 0·7). The other example of vectors at *B*, *B*′ and *B*″ (meeting at *C*) will be of interest in Chapter 11, 11.4.1.

lines (making a 45° angle; z = constant), in which the points of the lattice are those with integer coordinates (and the possible ones are those for which $t \geqslant 0$, $0 \leqslant z \leqslant t$[11]).

The behaviour of the gain $Y(n)$ (and of the individual outcomes) can be represented visually by means of its *path:* that is the jagged line joining the vertices $(n, Y(n))$ as in Figure 7.1a, where each 'step' upwards corresponds to a success, and each step downwards to a failure.[12] Each path of n initial steps on the lattice of Figure 7.2[13]

[11] In Figure 7.1b, if we take the two bisecting lines (with respect, that is, to the axes of Figure 7.1a), and therefore take as coordinates

$$\tfrac{1}{2}(t + y) = \text{number of successes } (= z, \text{ in Figure 7.1b})$$

and

$$\tfrac{1}{2}(t - y) = \text{number of failures } (= t - z, \text{ in Figure 7.1b}),$$

we have a system often used in other contexts (for example, in batch testing: a horizontal dash for a 'good' item, a vertical dash for a 'defective' item). This is convenient in this particular case, but has the disadvantage (a serious one if one wishes to study the random process) of not showing up clearly the independent variable (e.g. time), which one would like to represent along the horizontal axis.

[12] We observe, however, that this representation does not preserve the meaning of $Y(t)$, as given above, which requires it to change by a jump of ± 1 at the end of any interval $(n, n + 1)$ (and not linearly). The use of the jagged line is convenient, however, not only visually, and for the random-walk interpretation (see below, in the main text), but also for drawing attention to interesting features of the process. It is convenient, for example, to be able to say that one is *in the lead* or *behind* when the path is in the positive or negative half-plane, respectively (in other words, not according to the sign of $Y(n)$, which could be zero, but according to the sign of $Y(n) + Y(n + 1)$, where the two summands either have the same sign or one of them is zero).

[13] This representation also enables us to show clearly the probabilities at each step, and this is particularly useful when they vary from step to step (see, in particular, the end of Chapter 11, 11.4.1; the case of 'exchangeability'). All one has to do is the following: at each vertex, draw a vector, emanating from the vertex, and with components $(1, 2p - 1)$, the prevision vector of the next step (downwards or upwards; i.e. $(1, -1)$ or $(1, 1)$ with probabilities $1 - p$ and p).

(from 0 to a vertex on the nth vertical line) corresponds to one of the 2^n possible sequences of outcomes of the first n events.

The interpretation as a *random walk* is now immediate. The process consists in starting from the origin 0, and then walking along the lattice, deciding at each vertex whether to step upwards or downwards, the decision being made on the basis of the outcomes of the successive events. The same interpretation could, in fact, be made on the y-axis (each step being one up or one down), and this would be more direct, although less clear visually than the representation in the plane. When we think, in the context of 'random walk,' of Y_n as the distance of the moving point from the origin (on the positive or negative part of the y-axis) at time $t = n$, we are, in fact, using this representation: $Y_n = 0$ then corresponds to passage through the origin, and so on.

In fact, when one talks in terms of random walks, time is usually regarded as a parameter of the path (as for curves defined by parametric equations), so that, in general, we do not have an axis representing 'time' (and, if there is one, it is a waste of a dimension – even though visually useful). Normally, one uses the plane only for representing the random walk as a pair of (linearly independent) random functions of time (and the same holds in higher dimensions). An example of a random walk in two (or three) dimensions is given by considering the movement of a point whose coordinates at time n represent the gains of two (or three) individuals after n tosses (where, for example, each of them bets on Heads or Tails in any way he likes, with gains ±1).

We have mentioned several additional points which, strictly speaking, had little to do with our particular example, but will save us repeating ourselves when we come to less trivial situations. Moreover, it should be clear by now that the specific set-up we have considered will be suitable for dealing with any events whatsoever, no matter what their probabilities are, and no matter what the probability distributions of the random functions are.

7.3.3. If we restrict ourselves to considering the first n steps (events), the 2^n probabilities, non-negative, with sum 1, of the 2^n paths (i.e. of the 2^n products formed by sequences like $E_1\tilde{E}_2\tilde{E}_3E_4\ldots E_{n-1}\tilde{E}_n$) could be assigned in any way whatsoever. Thinking in terms of the random walk, the probabilities of the (n + l)th step being upward or downward will be proportional to the probabilities of the two paths obtained by making E_{n+1} or $\tilde{E}_{n+1}$ follow that determined by the n steps already made.

The image of probability as mass might also prove useful. The unit mass, initially placed at the origin 0, spreads out over the lattice, subdividing at each vertex in the manner we have just described (i.e., in general, depending on the vertex in question and the path travelled in order to reach it). One could think of the distribution of traffic over a number of routes which split into two forks after each step (provided the number of vehicles N is assumed large enough for one to be able to ignore the rounding errors which derive from considering multiples of $1/N$). The mass passing through an arbitrary vertex of the lattice, (t, y) say, comes from the two adjacent vertices on the left, $(t - 1, y - 1)$ and $(t - 1, y + 1)$ (unless there is only one, as happens on the boundaries $y = \pm t$) and proceeds by distributing itself between the two adjacent vertices on the right, $(t + 1, y - 1)$ and $(t + 1, y + 1)$. Thinking of the random walk as represented on the y-axis, all the particles are initially at the point 0 (or in the zero position), and after each time interval they move to one or other of the two adjacent points (or positions); from y to $y - 1$ or $y + 1$. Consequently, whatever the probabilities of such movements are, the mass is alternatively all in the even positions or all in the odd: in any case, we have some kind of *diffusion* process (and this is more than just an image). In many problems, it happens, for example, that under certain conditions the process comes to a halt, and this might correspond to the mass being stopped by coming up against an absorbing barrier.

7.3.4. *Further problems concerning Heads and Tails.* There are many interesting problems that we shall encounter where the probabilities involved are of a special, simple form. The most straightforward case is, of course, that in which the events E_n are independent; a further simplification results if they are equally probable. The simplest case of all is that of only two possible outcomes, each with probability $\frac{1}{2}$, and this is precisely the case of Heads and Tails that we have been considering.

The mass passing through a point *always divides itself equally* between the two adjacent points to the right. As a result of this, *each possible path of n steps always has the same probability*, $(\frac{1}{2})^n$, and *every problem concerning the probabilities of this process reduces to one of counting the favourable paths.*

Basing ourselves upon this simple fact, we can go back and give a systematic treatment of Heads and Tails, thinking of it now as a random process.[14] It is in this context that we shall again encounter well-known combinatorial ideas and results, this time in a form in which they are especially easy to remember, and which provides the most meaningful way of interpreting and representing them.

Prevision and standard deviation. For the case of Heads and Tails, the individual gains, $X_i = 2E_i - 1 = \pm 1$, are fair, and have unit standard deviations: in other words,

$$\mathbf{P}(X_i) = 0, \quad \sigma(X_i) = \mathbf{P}\left((\pm 1)^2\right) = \mathbf{P}(1) = 1.$$

The process itself is also fair, and the standard deviation of the gain in n tosses (which is, therefore, the quadratic prevision) is equal to $\surd n$: in other words,

$$\mathbf{P}(Y_n) = 0, \quad \sigma(Y_n) = \surd n.$$

The total number of successes is given by $S_n = \frac{1}{2}(n + Y_n)$ and so we have

$$\mathbf{P}(S_n) = \frac{1}{2}n, \quad \sigma(S_n) = \frac{1}{2}\surd n.$$

Successes in n tosses. We already know (case (A), Section 7.2.2) that the probability of h successes in n tosses is given by

$$\omega_h^{(n)} = \tbinom{n}{h}/2^n \quad (h = 0, 1, 2, \ldots, n). \tag{7.1}$$

[14] Figure 7.3 shows the results of successive subdivisions, $(\frac{1}{2},\frac{1}{2}),(\frac{1}{4},\frac{1}{2},\frac{1}{4}),(\frac{1}{8},\frac{3}{8},\frac{3}{8},\frac{1}{8})$ etc. Figure 7.4 shows a simple apparatus invented by Bittering. It is a box with two sets of divisions into compartments, one set being on top of the other and shifted through half the width of a compartment. The middle section of the bottom half of the box is filled with sand and then the box is turned upside down. The sand now divides itself between the two central compartments of what was the top half and is now the bottom. By repeatedly turning the box over (shaking it each time to ensure a uniform distribution of the sand within the compartments), one obtains successive subdivisions (the ones we have referred to above – those of Heads and Tails). By arranging the relative displacement of the overlapping compartments to be in the ratio p:1 – p, one can obtain any required Bernoulli distribution (see Section 7.4.2).

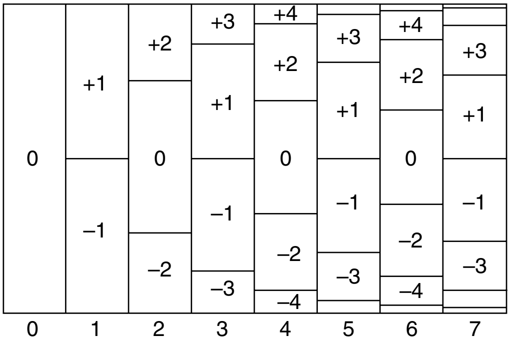

Figure 7.3 Subdivision of the probability for the game of Heads and Tails.

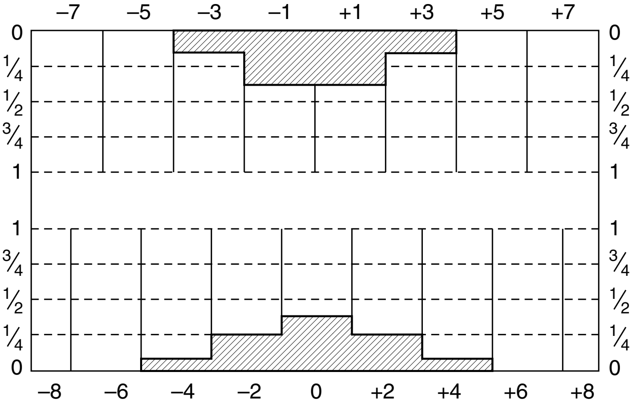

Figure 7.4 Bittering's apparatus. Probabilities of Heads and Tails: below, after four tosses; above, after three tosses.

This corresponds to the fact that $\binom{n}{h}$ is the number of paths which lead from the origin 0 to the point $[n, h]$.[15]

To see this, consider, at each point of the lattice, the number of paths coming from 0. This number is obtained by summing the numbers corresponding to the two points adjacent on the left, since all relevant paths must pass through one or other of these. 'Stiefel's identity', $\binom{n-1}{h-1}+\binom{n-1}{h}=\binom{n}{h}$, provides the key and leads one to the binomial coefficients of 'Pascal's triangle'. That the total number of paths is 2^n follows directly from the fact that at each step each path has precisely two possible continuations.

The identity which we mentioned in example (D) of Section 7.2.2 also finds an immediate application. Each of the $\binom{N}{H}$ jagged lines which lead to a given point $[N, H]$ must pass through the vertical at n at some point $[n, h]$, where, since n is less than N, $h = S_n$ must necessarily satisfy

$$H+n-N \leqslant h \leqslant H$$

[15] Recall that, in our notation, $[n, h]$ represents the fact that $S_{\mathrm{n}} = h$; in other words, it represents the point $(n, 2h - n)$, where $Y_{\mathrm{n}} = 2S_{\mathrm{n}} - n = 2h - n$.

(because, between n and N, there can be no reduction in either the number of successes or of failures). There are $\binom{n}{h}$ paths from the origin arriving at $[n, h]$, and $\binom{N-n}{H-h}$ paths leading from this point to $[N, H]$. There are, therefore, $\binom{n}{h}\binom{N-n}{H-h}$[16] paths from the origin to $[N, H]$ which pass through the given intermediate point $[n, h]$; summing over the appropriate values of h, we must obtain $\binom{N}{H}$, thus establishing the identity. Further, we see that

$$\binom{n}{h}\binom{N-n}{H-h}/\binom{N}{H}$$

is the probability of passing through the given intermediate point conditional on arriving at the given final destination (in other words, we obtain

$$\mathbf{P}\big(Y_n = 2h-n \,|\, Y_N = 2H-N\big) \quad \text{or} \quad \mathbf{P}(S_n = h \,|\, S_N = H);$$

we shall return to this later).

The rth *success*. Problem (F) of 7.2.4 can be tackled in a similar fashion, by reasoning in terms of crossings of sloping lines rather than vertical ones. In fact, the rth success is represented by the rth step upward; that is, the step which takes one from the line $y = 2r - 2 - t$ to the line $y = 2r - t$ (i.e. from the rth to the $(r + l)$th downward sloping line of the lattice, starting from $y = -t$). It is obvious that the rth failure can be dealt with by simply referring to upward sloping lines rather than downward ones. We have already shown the probability of the rth success at the hth toss to be $\binom{h-1}{r-1}/2^h$. We note that the favourable paths are those from 0 to $[h, r]$ whose final step is upward, that is passing through $[h - 1, r - 1]$, and that there are, in fact, $\binom{h-1}{r-1}$ of these.

If one is interested in considering the problem conditional on the path terminating at $[N, H]$, it is easily seen that there are $\binom{h-1}{r-1}\binom{N-h}{H-r}$ paths in which the rth success occurs at the hth toss (they must go from 0 to $[h - 1, r -1]$, and then, with a compulsory step, to $[h, r]$, and finally on to $[N, H]$). As above, we can sum over all possibilities to obtain $\binom{n}{h}$[17] (the sum being taken over $r \leqslant h \leqslant N - H + r$, since r successes cannot occur until at least r trials have been made, nor can there be more than $H - r$ failures in the final $H - r$ trials).

Dividing the sum by the total, we obtain, in this case also, the conditional probabilities (of the rth success at the hth toss, given that out of N tosses there are H successes; we must have $r \leqslant h \leqslant H N - H + r$). These are given by

[16] It is often not sufficiently emphasized that the *basic operation of combinatorial calculus is the product*; this should always be borne in mind, using this and many similar applications as examples.

[17] In this way, we arrive at meaningful interpretations of two well-known identities involving products of binomial coefficients:

$$\binom{N}{H} = \sum_{h=0\vee(H+n-N)}^{n\wedge H} \binom{n}{h}\binom{N-n}{H-h} \quad \text{(holding for each fixed } n\text{, with } 1 \leqslant n \leqslant N).$$

$$\binom{N}{H} = \sum_{h=r}^{N-H+r} \binom{h-1}{r-1}\binom{N-h}{H-r} \quad \text{(holding for each fixed } r\text{, with } 1 \leqslant r \leqslant H).$$

These simply give the number of paths from 0 to $[N, H]$, expressed in terms of the points at which they cross vertical (1st identity) or sloping (2nd identity) lines.

$$\binom{h-1}{r-1}\binom{N-h}{H-r}\Big/\binom{N}{H} = \mathbf{P}\left(S_{h-1}+1=S_h=r \mid S_N=H\right)$$
$$= \mathbf{P}\left(Y_{h-1}+1=Y_h=2r-h \mid Y_N=2H-N\right).$$

This result will also be referred to again later.

Gambler's ruin. The problem of the crossings of horizontal lines is more complicated, as, unlike the previous cases, more than one crossing is possible (in general, an unlimited number). It is, however, a very meaningful and important topic, and, in particular, relates to the classical gambler's ruin problem.

If a gambler has initial fortune c, then his ruin corresponds to his gain reaching $-c$. Similar considerations apply if two gamblers with limited fortunes play against each other. Here, we confine ourselves to just this brief comment but we note that, for this and for other similar problems (some of them important), arguments in terms of paths will turn out to be useful. In particular, we shall make use of appropriate *symmetries* of paths by means of Desiré André's celebrated *reflection principle* (in particular, Chapter 8 will deal with topics of this kind).

7.4 Some Particular Distributions

7.4.1. Before we actually begin our study of random processes, we shall, on the basis of our preliminary discussions, take the opportunity to examine a few simple problems in more detail, and to consider some particular distributions.

In order to avoid repetition later (and for greater effectiveness), we shall consider these distributions straightaway, both in the special forms that are appropriate for Heads and Tails, and in the more general forms. It should be noted that although the form of representation which we have adopted is a valid and useful one, the property of *fairness* (together with the principle of *reflection* and the *equal* probabilities of the paths) only holds for the special case of Heads and Tails.

In order to achieve some uniformity in notation, we shall always use X to denote the random quantity under consideration, and $p_h = \mathbf{P}(X = x_h)$ to denote the probability concentrated at the point x_h.[18] In the examples we shall consider, however, it turns out that the possible values of x_h are always integer (apart from changes in scale, $x_h = h$). For the particular case of the 'number of successes', we shall always use $\omega_h^{(n)} = \mathbf{P}(S_n = h)$ for the p_h.

7.4.2. *The Bernoulli* (*or binomial*) *distribution*. This is the distribution of $S_n = E_1 + E_2 + \ldots + E_n$ (or of $Y_n = 2S_n - n$, or of the frequency S_n/n – they are identical apart from an irrelevant change of scale) when the events E_h are *independent and have equal probabilities*, $\mathbf{P}(E_h) = p$. When $p = \frac{1}{2}$, as in the case of Heads and Tails, we have the *symmetric* Bernoulli distribution. The distributions are, of course, different for different n and p. Given n, the possible values are $x_h = h = 0, 1, 2,\ldots, n$ (or $x_h = a + hb = a, a + b, a + 2b,\ldots, a + nb$), and their probabilities are given by

$$p_h = \binom{n}{h} p^h \tilde{p}^{n-h} \quad \left(\text{if } p = \frac{1}{2}, p_h = \binom{n}{h}/2^n\right) \tag{7.2}$$

that is the $\omega_h^{(n)}$ of the process.

[18] We shall use *concentrated* rather than *adherent* (see Chapter 6) because, in these problems, the possible values can only be, by definition, the x_h themselves (finite in number, and, in any case, discrete).

For the case $p = \frac{1}{2}$, we know that $\mathbf{P}(X) = n/2$, and $\boldsymbol{\sigma}(X) = \surd n/2$ (see Section 7.3.4). Similarly, for arbitrary p, we see that $\mathbf{P}(X) = np$, $\boldsymbol{\sigma}(X) = \surd(np\tilde{p})$, because for each summand we have

$$\mathbf{P}(E_i) = \mathbf{P}(E_i^2) = p, \qquad \sigma^2(E_i) = \mathbf{P}(E_i^2) - \mathbf{P}^2(E_i) = p - p^2 = p\tilde{p}.$$

Hence, using the second-order properties, we obtain, without calculation,

$$\begin{aligned} &\sigma^2(X) = \sum_{h=0}^{n} \binom{n}{h} p^h \tilde{p}^{n-h} (h - np)^2 = np\tilde{p} \\ &\text{(in addition to } \mathbf{P}(X) = \sum_{h=0}^{n} \binom{n}{h} p^h \tilde{p}^{n-h} h = np). \end{aligned} \tag{7.3}$$

The behaviour of the $p_h = \omega_h^{(n)}$ in the case $p = \frac{1}{2}$ is governed by that of the binomial coefficients $\binom{n}{h}$. These are largest for central values ($h \simeq n/2$) and decrease rapidly as one moves away on either side. Unless one looks more carefully[19] at ratios like p_{h+1}/p_h, however, it is difficult to get an idea of *how rapidly* they die away:

$$\frac{\omega_{h+1}^{(n)}}{\omega_h^{(n)}} = \frac{n-h}{h+1} \quad \left(\text{in general, } \frac{n-h}{h+1} \cdot \frac{p}{\tilde{p}} \text{ for } p \neq \frac{1}{2}\right). \tag{7.4}$$

The same conclusions hold for general p, except that the maximum is attained for some $h \simeq np$ (instead of $\simeq \frac{1}{2} n$).

In fact, a consideration of Tchebychev's inequality suffices to show that the probability of obtaining values far away from the prevision is very small. Those h which differ from np by more than $n\varepsilon$ ($\varepsilon > 0$), that is corresponding to frequencies h/n not lying within $p \pm \varepsilon$, have, *in total*, a probability less than $\boldsymbol{\sigma}^2(X)/(n\varepsilon)^2 = np\tilde{p}/(n\varepsilon)^2 = p\tilde{p}/n\varepsilon^2$ (and this is far from being an accurate bound, as will be clear from the asymptotic evaluations which we shall come across shortly; equation 7.20 of Section 7.5.4).

Comments. The p_h can be obtained as the coefficients of the expansion of

$$(\tilde{p} + pt)^n = \sum_{h=0}^{n} p_h t^h = \sum_{h=0}^{n} \omega_h^{(n)} t^h$$

(the alternative notation being chosen to avoid any ambiguity in the discussion which follows). It suffices to observe that the random quantity

$$\prod_{i=1}^{n} (\tilde{E}_i + tE_i)$$

is the sum of the constituents multiplied by t^h, where h is the number of positive outcomes (giving, therefore, $S_n = h$). Its value is thus given by

$$\sum_{h=0}^{n} (S_n = h) t^h = t^{Sn},$$

[19] As is done, for the purpose of providing an elementary exposition, in B. de Finetti and F. Minisola, *La matematica per le applicazioni economiche*, Chapter 4. See also a brief comment later (in Section 7.6.3).

and its prevision, that is the characteristic function $\varphi(u)$ with $t = \mathrm{e}^{\mathrm{i}u}$ (or $u = -\mathrm{i}\log t$), by

$$\mathbf{P}\left(t^{Sn}\right) = \sum \mathbf{P}\left(S_n = h\right)t^h. \tag{7.5}$$

A generalization. In the same way, we observe that the result holds even if the E_i (always assumed stochastically independent) have different probabilities p_i. In this case, the $\omega_h^{(n)}$ are given by

$$\sum_{h=0}^{n} \omega_h^{(n)} t^h = \prod_{i=1}^{n} \left(\tilde{p}_i + p_i t\right). \tag{7.6}$$

On the other hand, this only expresses the obvious fact that $\omega_h^{(n)}$ is the sum of the products of h factors involving the p_i, and $n - h$ involving the complements $\tilde{p}_i$.

In particular, we see that $\sigma^2\left(S_n\right) = \sum_i p_i \tilde{p}_i$, and that this formula (like $np\tilde{p}$, of which it is an obvious generalization) continues to hold, even if we only have pairwise independence. (Recall that this is not sufficient for many other results concerning the distribution.)

7.4.3. *The hypergeometric distribution.* As in the previous case, we are interested in the distribution of $X = S_n$ (or $Y_n = 2S_n - n$, or S_n/n, which, as we have already remarked, only differ in scale). The difference is that we now *condition on the hypothesis that, for some given $N > n$, we have $S_N = H$.*

In deriving the required distribution, it suffices that the $\binom{N}{H}$ paths from 0 to $[N, H]$ appear equally probable to us. It does not matter, therefore, whether we choose to think in terms of Heads and Tails (where initially all 2^N paths were equally probable, and the paths compatible with the hypothesis remain such), or in terms of events which, prior to the hypothesis, were judged independent and equally probable, but with $p \neq \frac{1}{2}$ (because in the latter case all the remaining $\binom{N}{H}$ paths have the same probability, $p^H \tilde{p}^{N-H}$).

Instead of thinking in terms of these representations (whose main merit is that they show the links with what has gone before), it is useful to be able to refer to something rather more directly relevant. The following are suitable examples: drawings without replacement from an urn (containing N balls, H of which are white); counting votes (where a total of N have been cast, H of which are in favour of some given candidate); ordering N objects, H of which are of a given kind (N playing cards, H of which are 'Hearts'; N contestants, H of whom are female). In all these cases, the $N!$ possible permutations are all considered equally probable. (Or, at least, all $\binom{N}{H}$ possible ways of arranging the two different kinds of objects must be regarded as equally probable; it is these which correspond to the $\binom{N}{H}$ paths involving H upward steps and $N - H$ downward steps.)

Under the given assumptions (or information), each event $E_i (i \leqslant N)$ has probability $\mathbf{P}(E_i) = H/N$[20] (and, for convenience, we shall write $H/N = q$). These events are not independent; in fact, we shall see later that they are negatively correlated.

[20] Given that we assume the hypothesis $S_N = H$ to be already part of our knowledge or information, we take $\mathbf{P}(E)$ to mean $\mathbf{P}(E|S_N = H)$. In this situation, the E_i have probability $q = H/N$, but are not stochastically independent (even if they are the outcomes of coin tossing, or rolling a die etc., where, prior to the information about the frequency of successes out of N tosses, they were judged independent and equally probable). In particular, $p_h = \omega_h^{(n)} = \mathbf{P}(S_n = h)$ *in this case* is what we would have written as $\mathbf{P}(S_n = h|S_N = H)$ in the previous case.

Observe that, as a result of changes in the state of information, problems which were initially distinct may come to be regarded as identical, and assumptions about equal probabilities, or independence, may cease to hold (or conversely in some cases). These and other considerations will appear obvious to those who have entered into the spirit of our approach. Those who have come to believe (either through ignorance or misunderstanding) that properties like stochastic independence have an objective and absolute meaning that is inherent in the phenomena themselves, will undoubtedly find these things rather strange and mystifying.

The distribution that concerns us (for example, that of the number of white balls appearing in the first n drawings – or something equivalent in one of the other examples) will be different for every triple n, N and H (or, equivalently, n, N, q). For $q = \frac{1}{2}$, that is for $H = N - H = \frac{1}{2}N$, the distribution is symmetric; $\mathbf{P}(S_n = h) = \mathbf{P}(S_n = n - h)$. The possible values are the integers $x_h = h$, where $0 \vee H - (N - n) \leqslant h \leqslant n \wedge H$ (or $x_h = a + hb$; e.g. $= 2h - n$, or $= h/n$) and their probabilities are, as we saw already in 7.3.4,

$$
\begin{aligned}
p_h = \omega_h^{(n)} &= \frac{\binom{n}{h}\binom{N-n}{H-h}}{\binom{N}{H}} \\
&= \frac{\binom{H}{h}\binom{N-H}{n-h}}{\binom{H}{n}} \\
&= \binom{n}{h}\;\frac{\begin{array}{c}H(H-1)(H-2)\ldots(H-h+1)(N-H)(N-H-1)(N-H-2)\\ \ldots(N-H-(n-h)+1)\end{array}}{N(N-1)(N-2)\ldots(N-n+1)} \qquad (7.7)\\
&= \binom{n}{h} q h\tilde{q}^{n-h}\;\frac{\begin{array}{c}\left[\left(1-\frac{1}{H}\right)\left(1-\frac{2}{H}\right)\ldots\left(1-\frac{h-1}{H}\right)\right]\\ \left[\left(1-\frac{1}{N-H}\right)\left(1-\frac{2}{N-H}\right)\ldots\left(1-\frac{n-h-1}{N-H}\right)\right]\end{array}}{\left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)\ldots\left(1-\frac{n-1}{N}\right)}.
\end{aligned}
$$

The interpretation of the four different forms is as follows.

The *first form* (as we already know) enumerates the paths.

The *second form* enumerates those n-tuples, out of the total of $\binom{N}{H}$ that can be drawn from N events, which contain h out of the H, and $n - h$ out of the $N - H$.

The *third form* (which can be derived from the previous two) can be interpreted directly, observing that the probability of first obtaining h successes, and then $n - h$ failures, is given by the product of the ratios (of white balls, and then of black balls) remaining prior to each drawing:

$$
\frac{H}{N}\cdot\frac{H-1}{N-1}\cdot\frac{H-2}{N-2}\cdots\frac{H-h+1}{N-h+1}\cdot\frac{N-H}{N-h}\cdot\frac{N-H-1}{N-h-1}
\times\frac{N-H-2}{N-h-2}\cdots\frac{N-H-(n+h)+1}{N-n+1}.
$$

But this is also the probability for any other of the $\binom{n}{h}$ orders of drawing this number of successes and failures (even if the ratios at each drawing vary, the result merely involves permuting the factors in the numerator, and is, therefore, always the same).

We see already that, provided n is small in comparison to N, H and $N - H$, all the ratios differ little from q (the drawings already made do not seriously alter the composition of the urn). The results will, therefore, not differ much from the Bernoulli case (drawings from the same urn *with* replacement).

The *fourth form* shows the relation between the two cases explicitly, by displaying the correction factor.

The behaviour of the p_h in this case is similar to that of the Bernoulli case, and can be studied in the same way (by considering ratios p_{h+1}/p_h). The maximum is obtained by the largest h which does not exceed

$$nq\left[1-2/(N+2)\right]-(H-3)/(N+2)$$

(the reader should verify this!), and as one moves further and further away on each side, the p_h decrease. Compared with the Bernoulli case, the terms around the maximum are larger, and those far away are smaller. Some insight into this can be obtained by looking at the final formula.[21]

For the prevision, we have, of course, $\mathbf{P}(X) = nq = nH/N$. The standard deviation $\boldsymbol{\sigma}(X)$, on the other hand, is a little smaller than $\surd(np\tilde{p})$ (the result for the case of independence), and is given by

$$\sigma^2(X)=nq\tilde{q}\left[1-(n-1)/(N-1)\right].$$

In fact, if we evaluate the correlation coefficient r between two events ($r = r(E_i, E_j)$, $i \neq j$) we obtain $r = -1/(N - 1)$. More specifically,

$$\mathbf{P}(E_iE_j)=(H/N)\left((H-1)/(N-1)\right)=q^2(1-1/H)/(1-1/N),$$

from which it follows that

$$\begin{aligned} r &= \left[\mathbf{P}(E_iE_j)-\mathbf{P}(E_i)\mathbf{P}(E_j)\right]/\sigma(E_i)\sigma(E_j) \\ &= q^2\left[(N+H-1)/N(N-1)\right]/q\tilde{q} = -1/(N-1). \end{aligned}$$

[21] For this correction factor, the variant of Stirling's formula (equation 7.28) which is given in equation 30 (see 7.6.4) yields the approximation (for $n \ll N$)

$$\exp\left\{-\frac{1}{2}\frac{n}{N}\left[1+\frac{2\left(n-\frac{1}{2}\right)}{\eta(1-\eta)}(\xi-\eta)-\frac{n}{\eta(1-\eta)}(\xi-\eta)^2\right]\right\},$$

where we have set $\eta = H/N$ and $\xi = h/n$ (i.e. the percentage of white balls in the urn, and the frequency of white balls drawn in a sample of n, respectively). In the special case $\eta = \frac{1}{2}$ ($H = \frac{1}{2}N$; half the balls in the urn are white, half black), the expression simplifies considerably to give

$$\exp\left\{-\frac{1}{2}\frac{n}{N}\left[1-4n\left(\xi-\frac{1}{2}\right)^2\right]\right\}.$$

On the basis of this, we conclude that (approximately) the distribution gives higher probabilities than the Bernoulli distribution in the range where h lies between $\eta\eta \pm \surd[n\eta(1 - \eta)]$ (i.e. between $m \pm \sigma$), with a maximum at $n\eta$, and lower values outside this interval.

It is possible, however, to avoid the rather tiresome details (which we have skipped over) by using the argument already encountered in Chapter 4, 4.17.5, and observing that

$$\sigma^2(S_n) = nq\tilde{q} + 2\binom{n}{2}rq\tilde{q} = nq\tilde{q}\left[1+(n-1)r\right].$$

For $n = N$, we have $\boldsymbol{\sigma}(S_N) = 0$, because $(S_N = H)$ = the certain event, and hence

$$1+(N-1)r = 0, \qquad r = -1/(N-1). \tag{7.8}$$

Remarks. Note how useful it can be to bear in mind that apparently different problems may be identical, and how useful it can be to have derived different forms of expression for some given result, to have found their probabilistic interpretations, and to be in a position to recognize and use the simplest and most meaningful form in any given situation. Note, also, that one should be on the lookout for the possibility of reducing more complicated problems to simpler ones, both heuristically and, subsequently, by means of rigorous, detailed, analytical arguments, either exact or approximate.

In the case considered, we can go further and note that, by virtue of their interpretations within the problem itself, we have $\omega_h^{(n)} = \omega_{H-h}^{(N-n)}$. In other words, for given N and H, the distributions for complementary sample sizes n and $N - n$ are identical if we reverse $h = 0, 1,\dots, H$ to $H,\dots, 1, 0$ (and this can be seen immediately by glancing at the formula). It follows that, among other things, what is claimed to hold for 'small n must also hold for large n' (i.e. n close to N). The approximation does not work for central values ($n \sim \frac{1}{2}N$) and we note, in particular, that, for n lying between N and $N - H$, not all the values $h = 0, 1,\dots, n$ are possible (since they themselves must lie between H and n – (N – H).

7.4.4. *The Pascal distribution.* This is the distribution of X = 'the number of tosses required before the rth Head is obtained' (more generally, it arises for any independent events with arbitrary, constant probability p). Alternatively, it is the distribution of X such that $S_X = r > S_{X-1}$. By changing the scale, we could, of course, consider $X' = a + bX$. One example of this which often crops up is $X' = X - r$ = 'the number of failures preceding the rth success', but many of the other forms, such as those considered in the previous case, do not make sense in this context.

A new feature is that the distribution is unbounded, the possible values being $x_h = h = r, r + 1, r + 2,\dots$ (up to infinity, and, indeed, $+\infty$ must be included as a possible value, along with all the integers, since it corresponds to the case where the infinite set of trials result in less than r successes). In line with our previous policy, we shall avoid critical questions by deciding that if the rth success is not obtained within a maximum of N trials (where N is very large compared with the other numbers in question) we shall set $X = N$. (To be precise, we shall consider $X' = X \wedge N$ instead of X) Were we to consider $X' = X - r$, the possible values would be 0, 1, 2,… (and this is one of the reasons why this formulation is often preferred; another reason will be given in the discussion that follows; see equation 7.15).

For each r and p, we have, of course, a different distribution:

$$p_h = \binom{h-1}{r-1}p^r\tilde{p}^{h-r} \qquad (\text{if } p = \frac{1}{2}, p_h = \binom{h-1}{r-1}/2^h). \tag{7.9}$$

In fact (as we saw in (F), 7.2.4, for the special case $p = \frac{1}{2}$ in order to obtain $X = h$, we must have $r - 1$ successes in the first $h - 1$ trials, together with a success on the hth trial.

In terms of the random process representation, we are dealing with the crossing of the line $y = 2r - x$ (or, if one prefers, with the mass that ends up there if the line acts as an absorbing barrier).

Note that the series $\sum p_h$ sums up to 1. It must, of course, be convergent with sum $\leqslant 1$ by its very meaning; the fact that it is = 1, and not < 1, ensures that, as n increases, the probability that $X > n$ tends to zero (and, in particular, the probability that $X = \infty$ is 0).

So far as the behaviour of the distribution is concerned, the p_h again increase until they reach a maximum (attained for the greatest $h \leqslant r/p$), and then decrease to zero (asymptotically, like a geometric progression with ratio $\tilde{p}$). A more intuitive explanation of the increase is that it continues so long as the prevision of the number of successes, $\mathbf{P}(S_h) = hp$, does not exceed the required number of successes, r.

The geometric distribution. For $r = 1$, the Pascal distribution reduces to the special case of the geometric distribution:

$$p_h = p\tilde{p}^{h-r}, \tag{7.10}$$

forming a geometric progression ($p_1 = p$, with ratio $\tilde{p} = 1 - p$). If, for example, the first failure corresponds to elimination from a competition, this gives the probability of being eliminated at the hth trial, when the probability of failure at each trial is p. (N.B. For the purposes of this particular example, we have, for the time being, interchanged 'success' and 'failure'.) In particular, it gives the probability that a machine first goes wrong the hth time it is used, or that a radioactive atom disintegrates in h years time and so on, where the probability of occurrence is p on each separate occasion. (If the probability of death were assumed to be constant, rather than increasing with age, this would also apply to the death of an individual in h years time.)

The property of giving the same probability, irrespective of the passing of time, or of the outcomes of the phenomenon in the past, is known as the *lack of memory* property of the geometric distribution. The waiting time for a particular number to come up on the lottery[22] has, under the usual assumptions, a geometric distribution (the ratio is $\tilde{p} = \left(\frac{17}{18}\right) = 94 \cdot 44\%$, for a single city; $\tilde{p} = \left(\frac{17}{18}\right)^{10}$ = 56% for the whole set of ten cities). This provides further confirmation, if such were needed, of the absurdity of believing that numbers which have not come up for a long time are more likely to be drawn in future.

To put this more precisely: it is absurd to use the small probabilities of long waiting times, which are themselves evaluated on the basis of the usual assumptions, and are given by the geometric distribution (or to invoke their comparative rarity, statistically determined in accordance with it), to argue, on the basis of independence, against the very assumptions with which one started – that is the *lack of memory* property. If, on the other hand, someone arrived at a coherent evaluation of the probabilities by a different route, we might not judge him to be reasonable, but this would simply be a matter of opinion.

Finally, let us give the explicit expression for the case $r = 2$ (again, this could be thought of as elimination from a competition, but this time at the second failure): it reduces to $p_h = (h-1)p^2\tilde{p}^{h-2}$.

[22] *Translators' note.* See footnote 28 in Chapter 2.

Prevision and standard deviation. In order to calculate $\mathbf{P}(X)$ and $\boldsymbol{\sigma}(X)$, it turns out to be sufficient to do it for the case $r = 1$. We obtain

$$\mathbf{P}(X) = p\sum_{h=1}^{\infty} h\tilde{p}^{h-1} = p/(1-\tilde{p})^2 = 1/p, \tag{7.11}$$

$$\begin{aligned}\mathbf{P}(X^2) &= p\sum_{h=1}^{\infty} h^2\tilde{p}^{h-1} = p\tilde{p}\sum_{h=1}^{\infty} h(h-1)\tilde{p}^{h-2} + p\sum_{h=1}^{\infty} h\tilde{p}^{h-1} \\ &= 2p\tilde{p}/p^3 + 1/p = (2-p)/p^2\end{aligned} \tag{7.12}$$

(verify this!), and hence,

$$\sigma^2(X) = \mathbf{P}(X^2) - \mathbf{P}^2(X) = (1-p)/p^2.$$

For general r, it suffices to note that

$$\mathbf{P}(X) = r/p, \tag{7.13}$$

$$\sigma^2(X) = r(1-p)/p^2, \tag{7.14}$$

because (as we already observed for $\mathbf{P}(X)$ in the case $p = \frac{1}{2}$; see (*M*), 7.2.7) we can consider X as the sum of r terms, $X_1 + X_2 + \ldots + X_r$, stochastically independent, and each corresponding to $r = 1$ (X_i = the number of trials required after the $(i - 1)$th failure until the ith failure occurs).

Comments. This technique will also be useful in what follows: note that it can also be used for $X' = X - r$ if we consider r summands of the form $X_i' = X_i - 1$.

In this context (i.e. with h transformed into $h + r$), the p_h are given by

$$p_h = \binom{h+r-1}{r-1} p^r\tilde{p}^h = \binom{h\ +\ r-1}{h} p^r\tilde{p}^h = (-1)^h\binom{-r}{h} p^r\tilde{p}^h, \tag{7.15}$$

where the definition $\binom{x}{h} = x(x-1)\ldots(x-h+1)/h!$ is extended to cover any real x (not necessarily integer, not necessarily positive).

If we do this, the distribution then makes sense for any real $r > 0$. This generalized form of the Pascal distribution (which has integer r) is called the *negative binomial distribution* (simply because it involves the notation $\binom{-r}{h}$). For $r = 0$, the distribution is concentrated at the origin ($p_0 = 1, p_h = 0, h \neq 0$); for $r \sim 0$, we have $p_h \simeq r\tilde{p}^h/h$ ($h \neq 0$) (the logarithmic distribution; see Chapter 6, 6.11.2), and hence

$$p_0 \simeq 1 - r\sum_{h=1}^{\infty}\tilde{p}^h/h = 1 - r\log(1/p). \tag{7.16}$$

We shall make use of this later on, and the significance of the extension to noninteger r will also be explained.

The prevision in this case is clearly given by

$$\mathbf{P}(X') = \mathbf{P}(X) - r = r\tilde{p}/p, \tag{7.17}$$

whereas the standard deviation, $\sqrt{(r\tilde{p})}/p$, is unaltered; this also holds for noninteger r.

Another form. We have already seen (Section 7.3) that, if $S_N = H$ is assumed to be known, the problem of the location, h, of the rth success leads to the distribution

$$p_h = \binom{h-1}{r-1}\binom{N-h}{H-r}\Big/\binom{N}{H}, \tag{7.18}$$

rather than to the Pascal distribution. For an example where this distribution occurs, consider an election in which N votes have been cast and are counted one at a time. Suppose further that a candidate is to be declared elected (or a thesis accepted) when r votes in favour have been counted. Given that a total of $H \geqslant r$ out of N were actually favourable, equation 7.18 gives the probability that success is assured by the counting of the hth vote. (Another example, with $N = 90$ and $H = r = 15$, is given by the probability of completing a 'full house' at bingo with the hth number called.)

We shall restrict ourselves to considering the particularly simple case of $H = r = 1$, a case which is nonetheless important (and a study of the general case is left as an exercise for the reader). Clearly (even without going through the algebra), we have $p_h = 1/N$ for $h = 1, 2,\ldots, N$. If there has only been one success in N trials (or there is only one favourable vote in the ballot box, or only one white ball in the urn, or only one ball marked '90'), there is exactly the same probability of finding it on the first, second,..., or Nth (final) trial.

7.4.5. *The discrete uniform distribution.* This is the name given to the distribution of an X which can only take on a finite number of equally spaced possible values, each with the same probability: for example $x_h = h$, $h = 1, 2,\ldots, n$ (or $x_h = a + bh$), with all the $p_h = 1/n$. As examples, we have a fair die ($n = 6$), a roulette wheel ($n = 37$) or the game of bingo ($n = 90$).

It is easily seen that

$$\mathbf{P}(X)=\frac{1}{2}(n+1), \mathbf{P}(X^2)=(1^2+2^2+\ldots+n^2)/n=(4n^2+6n+2)/12,$$

from which (subtracting $\left[\frac{1}{2}(n+1)\right]^2=(3n^2+6n+3)/12$) we obtain

$$\sigma^2(X)=(n^2+1)/12, \quad \sigma(X)=(n/\sqrt{12})\sqrt{(1+1/n^2)}\simeq n/\sqrt{12}.$$

A random process (*Bayes–Laplace, Pólya*). Using this distribution as our starting point, we can develop a random process similar to that which led to the hypergeometric distribution. In fact, we consider successive drawings (without replacement) from an urn containing N balls, with the possible number of white balls being any of 0, 1, 2,..., N, each with probability l/(N +1). (This could arise, for example, if the urn were chosen from a set of N + 1 urns, ranging over all possible compositions, and there were no grounds for attributing different probabilities to the different possible choices.)

Let us assume, therefore, that $\omega_H^{(N)}=1/(N+1)(H=0,1,2\ldots,N)$, and that (as in the case of a known composition, H/N) all the permutations of the possible orders of drawing the balls are equally probable: in other words, that all the dispositions of H white and $N - H$ black balls (i.e. all the paths from 0 ending up at the same final point [N, H]) are equally probable. Each of these paths therefore has probability $1/\binom{N}{H}(N+1)$.

We shall now show that, under these conditions, the distribution for every S_n ($n < N$) is uniform, just as we assumed it to be for S_N: that is

$$\omega_h^{(n)} = 1/(n+1) \quad (h = 0, 1, 2, \ldots, n).$$

It can be verified, in a straightforward but tedious fashion, that

$$\sum_{H=0}^{N} \binom{n}{h}\binom{N-n}{H-n} \frac{1}{\binom{N}{H}(N+1)} = \frac{1}{n+1}$$

(the probabilities of the paths terminating at $[N, H]$ multiplied by the number of them that pass through $[n, h]$, the sum being taken over H). The proof by induction (from N to $N - 1$, $N - 2$,..., etc.) is much simpler, however, and more instructive. It will be sufficient to establish the step from N to $N - 1$. The probability $\omega_h^{(N-1)}$ that $S_{N-1} = h$ is obtained by observing that this can only take place if $H = h$ and the final ball is black, or if $H = h + 1$ and the final ball is white; each of these two hypotheses has probability $1/(N + 1)$ and the probabilities of a black ball under the first hypothesis and a white ball under the second are given by $(N - h)/N$ and $(h + 1)/N$, respectively. It follows that

$$\omega_h^{(N-1)} = \frac{1}{N+1}\left(\frac{N-h}{N} + \frac{h+1}{N}\right) = \frac{1}{N}. \tag{7.19}$$

Expressed in words: if all compositions are equally probable, so are all the frequencies at any intermediate stage. This property ($\omega_h^{(n)} = 1/(n+1)$) can also hold for all n (without them being bounded above by some pre-assigned N), and leads to the important Bayes–Laplace process (which we shall meet in Chapter 11, 11.4.3) or, with a different interpretation, to the Pólya process (Chapter 11, 11.4.4) with which it will be compared.

7.5 Laws of 'Large Numbers'

7.5.1. We now return to our study of the random process of Heads and Tails (as well as some rather less special cases) in order to carry out a preliminary investigation of what happens when we have 'a large number' of trials. This preliminary investigation will confine itself to qualitative aspects of the order of magnitude of the deviations. In a certain sense, it reduces to simple but important corollaries of an earlier result, which showed that *the order of magnitude increases as the square root of n* (the number of trials).

In the case of Heads and Tails $\left(p = \frac{1}{2}\right)$ the prevision of the *gain*, Y_n, is zero (the process is fair; $\mathbf{P}(Y_n) = 0$), and its standard deviation $\boldsymbol{\sigma}(Y_n)$ (which, in a certain sense, measures 'the order of magnitude' of $|Y_n|$) is equal to $\sqrt{n}$. The *number of successes* (Heads) is denoted by S_n and has prevision $\frac{1}{2}n$; its standard deviation (the order of magnitude, measured by σ) is equal to $\frac{1}{2}\sqrt{n}$. For the *frequency of successes*, S_n/n, the prevision and standard deviation are those we have just given, but now divided by n; that is $\frac{1}{2}$ and $\frac{1}{2}/\sqrt{n}$, respectively. In a similar way, one might be interested in the *average gain* (per toss), Y_n/n; this has prevision 0 and standard deviation $1/\sqrt{n}$.

The fact that

$$\mathbf{P}\left[\left(\frac{S_n}{n}-\frac{1}{2}\right)^2\right]=\frac{1}{4n}\to 0$$

is expressed by saying that *the frequency converges in mean-square to* $\frac{1}{2}$. This implies (see Chapter 6, 6.8.3) that it also converges *in probability*. Similarly, the average gain converges to 0 (both in mean-square and in probability).

We recall that convergence in probability means that, for positive ε and θ (however small), we have, for all n greater than some N,

$$\mathbf{P}\left(\left|\frac{S_n}{n}-\frac{1}{2}\right|>\varepsilon\right)<\theta.$$

A straightforward application of Tchebychev's inequality shows that the probability in our case is less than $\sigma^2/\varepsilon^2=1/4n\varepsilon^2$.

7.5.2. When referring to this result, one usually says, in an informal manner, that after a large number of trials it is *practically certain* that the frequency becomes *practically equal* to the probability. Alternatively, one might say that 'the fluctuations tend to cancel one another out.' One should be careful, however, to avoid exaggerated and manifestly absurd interpretations of this result (a common trap for the unwary). Do not imagine, for example, that convergence to the probability is to be expected because future discrepancies should occur in such a way as to 'compensate' for present discrepancies by being in the opposite direction. Nor should one imagine that this holds for the absolute deviations. It is less risky to gamble just a few times (e.g. ten plays at Heads and Tails at 1000 lire a time) than it is to repeat the same bet many times (e.g. a 1000 plays are 10 times more risky; 10 = √(1000/10)). On the other hand, it would be less risky to bet 1000 times at 10 lire a time. Furthermore, if at a certain stage one is losing – let us say 7200 lire – the law of large numbers provides no grounds for supposing that one will 'get one's own back.'[23] In terms of prevision, this loss remains forever at the same level, 7200 lire. The future gain (positive or negative) has prevision zero but, as one proceeds, the order of magnitude becomes larger and larger and, eventually, it makes the loss already suffered *appear negligible*. It is in this sense, and only in this sense, that the word 'compensate' might reasonably be used, since one would then avoid the misleading impression that it usually conveys. The fact remains, however, that the loss has already been incurred.

Observe once again how absurd it would be to imagine, a priori, some sort of correlation – which would be a consequence of laws and results derived on the basis of an assumption of independence!

7.5.3. In addition to this, one should note that the property we have established concerns the probability of a deviation $>\varepsilon$ between the probability and the frequency for

[23] The illusory nature and pernicious influence of such assumptions are referred to in a popular, witty saying (possibly Sicilian in origin), in itself rather remarkable, given that popular prejudice seems on the whole to incline towards the opposite point of view. The saying concerns the answer given by a woman to a friend, who has asked whether it was true that her son had lost a large amount of money gambling: '*Yes, it's true*', she replies, '*But that's nothing: what is worse is that he wants to get his own back*!'.

an individual n (although it can be for any $n \geqslant N$). This clearly does not imply – although the fact that it does not can easily escape one's notice – that the probability of an 'exceptional' deviation occurring at least once for an n between some N and an $N + K$ greater than N is also small.

It is especially easy to overlook this if one gets into the habit of referring to events with small probability as 'impossible' (and even worse if one appears to legitimize this bad habit by giving it a name – like 'Cournot's principle', Chapter 5, 5.10.9). In fact, if an 'exception' were impossible for every individual case $n \geqslant N$, it would certainly be impossible to have even a single exception among the infinite number of cases from $n = N$ onwards.

If one wanted to use the word 'impossible' in this context without running into these problems, it would be necessary to spell out the fact that it should not be understood as meaning 'impossible', but rather 'very improbable'. However, anyone who states that 'horses are potatoes', making it clear that when it refers to horses the meaning of 'potato' is not really that of 'potato' (but rather that of 'horse'), would probably do better not to create useless terminological complications in the first place (since, in order for it not to be misleading, it must be immediately followed by a qualification which takes away its meaning).

Now let us return to the topic of convergence. If the probability of a deviation $|S_n/n - \frac{1}{2}|$ at Heads and Tails being greater than ε were actually equal to $1/4n\varepsilon^2$, then, for any N, in some interval from N to a sufficiently large $N + K$ the prevision of the number of 'exceptions' (deviations $>\varepsilon$) would be arbitrarily large (approximately $(l/4\varepsilon^2)\log(1 + K/N)$). This follows from the fact that the series $\sum 1/n$ is divergent and the sum between N and $N + K$ is approximately equal to $\log(l + K/N)$. In fact, as we shall see later, the result we referred to at the beginning of 7.5.3 does hold. It just so happens that the Tchebychev inequality, although very powerful in relation to its simplicity, is not sufficient for this more delicate result. Stated mathematically, we have, for arbitrary positive ε and θ,

$$\mathbf{P}\left(\max_{N \leqslant n \leqslant N+K}\left|\frac{S_n}{n} - \frac{1}{2}\right| > \varepsilon\right) < \theta,$$

provided N is sufficiently large (K is arbitrary).[24] This form of stochastic convergence is referred to as *strong convergence* and the result is known as the '*strong law of large numbers*'. By way of contrast, the word 'strong' is replaced by '*weak*' when we are referring to convergence in probability, or to the previous form of the law of large numbers.

7.5.4. In order to fix ideas, we have referred throughout to the case of Heads and Tails. Of course, the results also hold for $p \neq \frac{1}{2}$ (except that we then have to write $\sigma^2 = p\tilde{p}$, which is $< \frac{1}{4}$ unless $p = \frac{1}{2}$) and even in the case where the $p_i = \mathbf{P}(E_i)$ vary from event to event (provided $\sum p_i \tilde{p}_i$ diverges, which may not happen if the p_i get too close to the extreme values 0 and 1). In the latter case, our statement would, in general, assert that the difference between the frequency S_n/n in the first n trials and the arithmetic mean of the probabilities, $(p_1 + p_2 + \ldots + p_n)/n$, tends to zero (in mean-square and in probability). Only if the arithmetic mean tends to a limit p (or, as analysts would say, if the p_i are a

[24] Were it not for our finitistic scruples (see Chapter 6 and elsewhere), we could do as most people do, and write $\sup(n \geqslant N)$ in place of $\max(N \leqslant n \leqslant N + K)$, saying that it is 'almost certain' (i.e. the probability = 1) that $\lim(\mathrm{S_n}/n) = \frac{1}{2}$ (in the sense given).

sequence converging to *p in the Cesàro sense*) does the previous statement in terms of deviation from a fixed value hold (and the fixed value would then be the limit p).

We can, however, say a great deal more on the basis of what we have established so far. The only properties we have made use of are those of the previsions and standard deviations of the gains of individual bets, $2E_i - 1$, and of their sums, Y_n. It is easy to convince oneself that for the conclusion (weak convergence) to hold we only require that the gains X_i ($i = 1, 2,\ldots$) have certain properties. For example, it suffices that they have zero prevision and are (*pairwise*) *independent with constant, finite standard deviations.* More generally, we only require that they are (pairwise) uncorrected and that the standard deviations $\boldsymbol{\sigma}(X_i) = \sigma_i$ are bounded, and such that $\sum\sigma_i^2$ diverges. Considering the case of zero prevision for convenience, we have

$$Y_n/n = \left(X_1 + X_2 + \ldots + X_n\right)/n \overset{\cdot}{\to} 0 \quad \left(\text{and hence} \overset{<}{\to} 0\right).$$

Expressed in words: the (weak) law of large numbers holds for sums of uncorrelated random quantities under very general conditions, in the sense that *the arithmetic mean, Y_n/n, tends to* 0 *in quadratic prevision, and the probability of its having an absolute value* $>\varepsilon$ (*an arbitrary, preassigned positive value*) *also tends to* 0.

The *strong law of large numbers* also holds under very general conditions. The argument which ensures its validity if the sum of the probabilities p_h of 'exceptions' (deviations $|Y_h/h| > \varepsilon$) converges, turns out to be sufficient if these probabilities are evaluated on the basis of the normal distribution, and this will be the case if the X_h are assumed to be independent with standard unit normal distributions ($m = 0$, $\sigma = 1$). Asymptotically, however, this property also holds in the case of Heads and Tails, and for any other X_h which are identically distributed with finite variances (let us assume $\sigma = 1$).[25] We shall, as we mentioned above, restrict ourselves to the proof based on the convergence of Σp_h. Afterwards, we shall mention the possibility of modifications which make the procedure much more powerful.

Since the distribution function of the (standard unit) normal cannot be expressed in a closed form, it is necessary, in problems of this kind, to have recourse to an asymptotic formula (which can easily be verified – by L'Hospital's rule, for example):

$$1 - F(x) = \frac{1}{\sqrt{(2\pi)}}\int_x^\infty e^{-\frac{1}{2}z^2}\,\mathrm{d}z \sim \frac{1}{\sqrt{(2\pi)}x}e^{-\frac{1}{2}x^2} \quad \left(\text{as}\, x \to +\infty\right). \tag{7.20}$$

It follows, since Y_h has standard deviation $\sqrt{h}$, that $|Y_h/h| > \varepsilon$ can be thought of as $|Y_h/\sqrt{h}| > \varepsilon\sqrt{h} = \varepsilon\sqrt{h}$ × the standard deviation of the standardized distribution, and, therefore,

$$\begin{aligned} p_h &= 1 - F\left(\varepsilon\sqrt{h}\right) + F\left(-\varepsilon\sqrt{h}\right) = 2\left[1 - F\left(\varepsilon\sqrt{h}\right)\right] \\ &\sim \frac{2}{\sqrt{(2\pi)}\varepsilon\sqrt{h}}e^{-\frac{1}{2}h\varepsilon^2} = \frac{K}{\sqrt{h}}e^{-ch}. \end{aligned}$$

[25] That the normal distribution frequently turns up is a fact which is well known, even to the layman (though the explanation is often not properly understood). The case we are referring to here will be dealt with in Section 7.6; we shall not, therefore, enter into any detailed discussion at present.

The geometric series $\sum e^{-ch}$ converges, however, and hence, *a fortiori*, so does the series $\sum p_h$, with the terms divided by $\surd h$. The remainder, from some appropriate N onwards, is less than some preassigned θ, and this implies that the probability of even a single 'exceptional deviation', $|Y_h/h| > \varepsilon$, for h lying between N and an arbitrary $N + M$, is less than θ. (If countable additivity were admitted, one would simply state the result 'for all $h \geqslant N$'.)

The conclusion can easily be strengthened by observing that convergence still holds even if the constant ε is replaced by some $\varepsilon(h)$ decreasing with h; for example,

$$\varepsilon(h) = \surd(2a\ \log h)/\surd h, \quad \text{with} \quad a > 1.$$

We then have $h\varepsilon^2 = 2a \log h$, and

$$p_h = \mathbf{P}\big(|Y_h / h\varepsilon| > (h)\big) = \Big[K / \surd\big(2a\ \log\ h\big)\Big]\ \mathrm{e}^{-a \log h} = (\ldots)h^{-a}.$$

But the terms (...) tend to zero, the series $\sum h^{-a}$ ($a > 1$) converges, and, *a fortiori*, $\sum p_h$ converges. Expressed informally, this implies that, from some N on, it is 'almost certain that Y_h will remain between $\pm c\surd(2h \log h)$' for $c > 1$.

The argument that follows exemplifies the methods that could be used to further strengthen the conclusions. Indeed, we shall see precisely how it is that one arrives at a conclusion which is, in a certain sense, the best possible ('we shall see', in the sense that we will sketch an outline of the proof without giving the details).

We note that were we to consider only the possible exceptions (Y_h lying outside the interval given above) at the points $h = 2^k$, instead of at each h, we could obtain the same convergent series by taking

$$\varepsilon(h)\surd h = \surd\big(2a\ \log\ k\big) \sim \surd\big(2a\ \log\ \log\ h\big), \quad \text{instead of} \surd\big(2a\ \log\ h\big).$$

A conclusion that only applies to the values $h = 2^k$ is, of course, of little interest, but it is intuitively obvious that we certainly do not require a check on all the h. The graph of $y = Y(h)$ can scarcely go beyond the preassigned bounds if one checks that it has remained within them by scanning a sufficiently dense sequence of 'check points'. Well, one can show that the check points $h = 2^k$ (for example) are sufficiently dense for one to be able to conclude that – again expressed informally – it is almost certain that all the Y_h from some N on (an N which cannot be made precise), will, in fact, remain within much smaller bounds of the form $\pm c\surd(2h \log \log h)$, for $c > 1$.

What makes this result important is that, conversely, if $c < 1$, it is 'practically certain that these bounds will be exceeded, however far one continues'. This is the celebrated *law of the iterated logarithm*, due to Khintchin.

Note that, in order to prove the converse which we have just stated, the divergence of the series is not sufficient, unless the events are independent (Borel–Cantelli lemma). In the case under consideration, we do not have independence. We do, however, have the possibility of reducing ourselves to the latter case, because, if h'' is much larger than h' the contribution of the increment between h' and h'' (which is independent of $Y(h')$) is the dominating term in $Y(h'') = Y(h') + [Y(h'') - Y(h')]$.

All these problems can be viewed in a more intuitive light (and can be dealt with using other techniques, developed on the basis of other approaches) if we base ourselves on random processes on the real line (and, so far as the results we have just mentioned are concerned, in particular on the Wiener–Lévy process). It will be a question of studying the graph $y = Y(t)$ of a random function in relation to regions like $|y| \leqslant y(t)$ (a preassigned function), by studying the probability of the graph entering or leaving the region, either once, or several times, or indefinitely.

Finally, let us just mention the standard set of conditions which are sufficient for the validity of the strong law. The X_h are required to be independent, and such that $\sum\sigma_h^2 / h^2$ converges (the Kolmogorov condition). The proof, which is based on an inequality due to Kolmogorov, one which, in a certain sense, strengthens the Tchebychev inequality and on the truncation of the 'large values' of the X_h, goes beyond what we wanted to mention at this stage.

In the classical case (that of independent events with equal probabilities), the weak and strong laws of large numbers are also known as the Bernoulli and Cantelli laws, respectively.

7.5.5. *The meaning and value of such 'laws'.* In addition to their intrinsic meaning, both mathematically and probabilistically, the laws of large numbers, and other asymptotic results of this kind, are often assigned fundamental rôles in relation to questions concerning the foundations of statistics and the calculus of probability. It seems appropriate to provide some discussion of this fact, both in order to clarify the various positions, and, in particular, to clarify our own attitude.

For those who seek to connect the notion of probability with that of frequency, results which relate probability and frequency in some way (and especially those results like the 'laws of large numbers') play a pivotal rôle, providing support for the approach and for the identification of the concepts. Logically speaking, however, one cannot escape from the dilemma posed by the fact that the same thing cannot both be assumed first as a definition and then proved as a theorem; nor can one avoid the contradiction that arises from a definition which would assume as certain something that the theorem only states to be very probable. In general, this point is accepted, even by those who support a statistical-frequency concept of probability; the attempts to get around it usually take the form of singling out, separating off, and generally complicating, particular concepts and models.

An example of this is provided by the 'empirical law of chance'. A phrase created for the purpose of affirming the *actual occurrence* of something the 'law of large numbers' states to be *very probable* comes to be presented as an *experimental fact*. Another example is provided by 'Cournot's principle': this states, as we mentioned in Chapter 5, 5.10.9, that 'an event of small probability does not occur', and covers the above, implicitly, as a special case. Sometimes, the qualification 'never, or almost never' is added, but although this removes the absurdity, in doing so it also takes away any value that the original statement may have had.

In any case, this kind of thing does nothing to break the vicious circle. It only succeeds in moving it somewhere else, or disguising it, or hiding it. A veritable labour of Sisyphus! It always ends up as a struggle against *irresolvable* difficulties, which, in a well-chosen phrase of B.O. Koopman, '*always retreat but are never finally defeated*, unlike Napolean's Guard'.

In order for the results concerning frequencies to make sense, it is necessary that the concept of probability, and the concepts deriving from it that appear in the statements and proofs of these results, should have been defined and given a meaning beforehand.

In particular, a result that depends on certain events being uncorrelated, or having equal probabilities, does not make sense unless one has defined in advance what one means by the probabilities of the individual events. This requires that probabilities are attributed to each of the given events (or 'trials'), that these all turn out to be equal and that, in addition, probabilities are attributed to the products of pairs of events, such that these are all equal and, moreover, equal to the square of the individual probabilities. In using the word 'attributed', we have, of course, used a word which fits in well with the subjectivistic point of view; in this context, however, it would make no difference if we were to think of such probabilities as 'existing', in accordance with the 'logical' or 'necessary' conception. In fact, the criticisms of the frequentistic interpretation made by Jeffreys, for instance, and the case against it which he puts forward (closely argued, and, I would say, unanswerable[26]), are in complete accord with the views we have outlined above. We acknowledge, of course, that there are differences between the necessary and subjectivistic positions (the latter denies that there are logical grounds for picking out *one single* evaluation of probability as being *objectively* special and 'correct'), but we regard this as of secondary importance in comparison with the differences that exist between, on the one hand, conceptions in which probability is probability (and frequency is just one of the ingredients of the 'outside world' which might or might not influence the evaluation of a probability) and, on the other hand, conceptions in which probability is, to a greater or lesser extent, a derivative of frequency, or is an idealization or imitation of it.

7.5.6. From our point of view, the law of large numbers forms yet another link in the chain of properties which justify our making use of expected or observed frequencies in our (necessarily subjective) evaluations of probability. We now see how to make use of the prevision of a frequency in this connection. The law of large numbers says that, *under certain conditions*, the value of the probability is *not only equal to the prevision* $\mathbf{P}(X)$ of a frequency X, but, moreover, *we are almost certain that X will be very close to this value* (getting ever closer, in a way that can be made precise, as one thinks of an even larger number of events).

This really completes the picture for the special case we have considered. Rather than introducing new elements into the situation (something we shall come across when we deal with exchangeable events in Chapter 11, and in similar contexts), we shall use these results in order to consider rather more carefully the nature of this special case: that is independent events with equal probabilities. It is important to realize that these assumptions, so apparently innocuous and easily accepted, contain unsuspected implications. To judge a coin to be 'perfectly fair so far as a single toss is concerned', means that one considers the two sides to be equally probable on this (the first) toss, or on any other toss for which one does not know the outcomes of the previous tosses. To judge a coin to be 'perfectly fair so far as the random process of Heads and Tails is concerned' is a very different matter.

The latter is, in fact, an extremely rash judgement that commits one to a great deal. It *commits* one, for example, to evaluating the probabilities as $\frac{1}{2}$ at each toss, *even if all*

[26] See. Harold Jeffreys, *Scientific Inference*, Cambridge University Press; 1st edn (1931), 2nd edn (1957) and *Theory of Probability*, Oxford University Press; 1st edn (1938), 2nd edn (1948), 3rd edn (1961). Particularly relevant are the following: Section 9.21 of the first work, entitled 'The frequency theories of probability', and Sections 7.03–7.05 of the second, in Chapter 7, 'Frequency definitions and direct methods'.

the previous tosses (a thousand, or a million, or 10^{1000},...) *were all Heads*, or Heads and Tails alternately, and so on. Another consequence (although one which is well beyond the range of intuition) is that, for a sufficiently large number of tosses, one considers it advantageous to bet on the frequency lying between 0·49999 and 0·50001 rather than elsewhere in the interval [0, 1] (and the same holds for $0{\cdot}5 \pm 10^{-1000}$ etc.). I once remarked that '*the main practical application of the law of large numbers consists of persuading people how unrealistic and unreasonable it is, in practice, to make rigid assumptions of stochastic independence and equal probabilities*'. The remark was taken up by L.J. Savage, to whom it was made, and given publicity in one of his papers. It was intended to be witty, in part facetious and paradoxical, but I think that it is basically an accurate observation.

Notwithstanding its great mathematical interest, there is clearly even less to be said from a realistic and meaningful point of view concerning the strong law of large numbers.

7.5.7. *Explanations based on 'homogeneity'.* First of all, it is necessary to draw attention to the upside-down nature of the very definitions of notions (or would-be notions) like those of *homogeneous* events, *perfect* coins and so on. Any definition that is framed in objective, physical terms, or whatever, is not suitable, because it cannot be used to *prove* that a given probabilistic opinion is a logical truth, nor can it *justify its imposition* as an article of faith.

If one wants to make use of these, or similar, notions, it is clear, therefore, that their meanings can only come about and be made precise as expressions of particular instances of probabilistic opinions (opinions which, had one already attributed to these notions a metaphysical meaning, preceding these personal opinions, one would have called a *consequence* of it).

I recall a remark, dating from about the time of my graduation, which has remained engraved upon my memory, having struck me at the time as being very accurate. A friend of mine used to say, half-jokingly, and in a friendly, mocking way, that it was never enough for me to define a concept, but that I needed to 'definettine' it. In actual fact, I had, by and large, adopted the mode of thinking advocated by authors like Vailati and Calderoni (or perhaps it would be more accurate to say that I found their approach to be close to my own). Papini used to say of Calderoni that 'what he wanted to do was to show what precautions one ought to take, and what procedures one ought to use, in order to arrive at statements which make sense'.[27] On the other hand, it was precisely this form of reasoning which, in successive waves, from Galileo to Einstein, from Heisenberg to Born, freed physics – and with it the whole of science and human thought – from those superstructures of absurd metaphysical dross which had condemned it to an endless round of quibbling about pretentious vacuities.

At the same time, and for the reasons we have just given, any attempt to define a coin as 'perfect' on the basis of there being no objective characteristics that prevent the probability of Heads from being $p = \frac{1}{2}$, or different tosses from being stochastically independent, is simply a rather tortuous way of making it appear that the above-mentioned objective circumstances play a decisive rôle. In fact, they are mere window dressing. The real meaning only becomes clear when these circumstances are pushed on one side and one simply proceeds as follows (and, in doing so, discovers that there are two possible meanings of 'perfect'): we shall use the expression *perfect coin* in the *weak* sense as a shorthand statement of the fact that we attribute equal probabilities $\left(\frac{1}{2}\right)$ to each of the

[27] G. Papini, *Stroncature*, No. 14: 'Mario Calderoni'; G. Vailati, *Scritti* (in particular, see those works quoted in the footnotes to Chapter 11, 11.1.5).

two possible outcomes of a toss; we use the expression in the *strong* sense if we attribute equal probabilities $\left(\left(\frac{1}{2}\right)^n\right)$ to each of the 2^n possible outcomes of n tosses, for any n. This does not mean, of course, that in making such a judgment it is not appropriate (or, even less, that it is not admissible) to take into account all those objective circumstances that one considers relevant to the evaluation of probability. It merely implies that the evaluation (or, equivalently, the identification and listing of the circumstances that might 'reasonably' influence it) is not a matter for the theory itself, but for the individual applying it. From his knowledge of the theory, the individual will have at his disposal various auxiliary devices to aid him in sharpening his subjective analysis of individual cases; the standard schemes will serve as reference points for his idealized schemes. There is no way, however, in which the individual can avoid the burden of responsibility for his own evaluations. The key cannot be found that will unlock the enchanted garden wherein, among the fairy rings and the shrubs of magic wands, beneath the trees laden with monads and noumena, blossom forth the flowers of *Probabilitas realis*. With these fabulous blooms safely in our button-holes we would be spared the necessity of forming opinions, and the heavy loads we bear upon our necks would be rendered superfluous once and for all.

7.5.8. Having dealt with the logical aspect, it remains to consider, in a more detailed fashion, the criticisms of those discussions based upon *homogeneity*, both from a practical point of view and from the point of view of the 'realism' of such a notion in relation to actual applications. It is curious to observe that these kinds of properties (independence with equal probabilities) are even less realistic than usual in precisely those cases that correspond to the very empirical–statistical interpretation which claims to be the most 'realistic' (i.e. those attributing the 'stability of the frequency' to quasi-'physical' peculiarities of some phenomenon possessing 'statistical regularity').

Can we really believe that a coin – 'perfect' so far as we can see – provides the perfect example of a phenomenon possessing these 'virtues'? There appears to be room for doubt. Is it not indeed likely that 'suspicious' outcomes would lead us to re-evaluate the probability, somehow doubting its perfection, or the manner of tossing, or something else?

By way of contrast, we would have less reason for such suspicions and doubts if, from time to time, or even at each toss, the coin were changed. This would be even more true if coins of different denomination were used and the person doing the tossing were replaced, and more so again if the successive events considered were completely different in kind (for example: whether we get an even or odd number with a die, or with two dice, or in drawing a number at bingo, or for the number plate of the first car passing by, or in the decimal place of the maximum temperature recorded today and so on; whether or not the second number if greater than the first when we consider number plates of cars passing by, ages of passers-by, telephone numbers of those who call us up and so on; it is open to anyone to display their imagination by inventing other examples). Under these circumstances, it seems very unlikely that a 'suspicious' outcome, whatever it was, would lead one to expect similar strange behaviour from future events, which lack any similarity or connection with those that have gone before.[28]

[28] We have used the qualification 'suspicious' only after careful consideration (we have avoided, for example, 'exceptional', or 'strange', or 'unlikely'). Now is not the appropriate time for a detailed examination of this question, however. This will come later (in Chapter 11, 11.3.1), and will clarify the meaning of this term and the reasons why it was chosen.

This demonstrates that the homogeneity of the events (the fact of their being, in some sense, 'trials of the same phenomenon', endowed with would-be statistical virtues of a special kind) is by no means a necessary prerequisite for the possible acceptance of the properties of independence with equal probabilities. In fact, it is a positive obstacle to such an acceptance. If, in such a case, the above properties are accepted, it is not that they should be thought of as valid *because of homogeneity* but, if at all, *in spite of homogeneity* (and it is much easier to accept them in other cases *because of heterogeneity*).[29] Despite all this, we continue to hear the exact opposite, repeated over and over, with the tiresome insistence of a silly catchphrase.

The 'laws of chance' (although it is rather misleading to refer to them in this way) express, instead, precisely that which one can expect from a maximum of disorder, in which any kind of useful knowledge is lacking. Any increase in one's knowledge of the phenomena and of their 'properties' would, if it were to be at all useful, lead one to favour some subset of the 2^n possible outcomes, and hence would lead to an evaluation of probabilities which are *better in this specific case* (with respect to the judgment of the individual who makes his evaluation after taking it into account) than those which would be valid in the absence of any information of this kind. There exists no information, knowledge, or property, that can strengthen or give 'physical' (or philosophical, or any other) meaning to the situation which corresponds to a perfect symmetry of ignorance.[30]

7.6 The 'Central Limit Theorem'; The Normal Distribution

7.6.1. If one draws the histograms of the distribution of Heads and Tails (the binomial distribution with $p = \frac{1}{2}$) and compares them for various values of n (the number of tosses), one sees that the shape remains the same (apart from discontinuities and truncation of the tails, features which arise because of the discreteness, and tend to vanish as n increases). The shape, in fact, suggests that one is dealing with the familiar normal distribution (the Gaussian distribution, or 'distribution of errors', which we mentioned briefly in Chapter 6, 6.11.3, and will further treat in Section 7.6.6; see Figure 7.6). Figure 7.5 gives the histogram for $n = 9$ (which is, in fact, a very small n!), together with the density curve. The agreement is already quite good, and the curve and the boundary of the histogram would rapidly become indistinguishable if we took a larger n (not necessarily very large).

In order to adjust the histograms to arrive at a unique curve, it is, of course, necessary to make an appropriate change of scale (we are concerned with convergence to a *type* of distribution; see Chapter 6, 6.7.1). The standard procedure of transforming to $m = 0$ and $\sigma = 1$ (Chapter 6, 6.6.6) is convenient and this is what we have done in the figure.[31]

[29] A fuller account of this may be found in 'Sulla "compensazione" tra rischi eterogenei', *Giorn. Ist. Ital. Attuari* (1954), 1–21.

[30] The following point has been made many times, and should be unnecessary. We are not speaking of exterior symmetries (which could exist), nor of 'perfect ignorance' (which cannot exist – otherwise, we would not even know what we were talking about), but about symmetry of judgment as made by the individual (in relation to the notion of indifference which he had prior to obtaining information, whether a great deal or only a small amount).

[31] This holds for the actual distribution (discrete: the mass of every small rectangle concentrated at the centre). If one thinks of it as diffused, one must modify this slightly (an increase) as we shall see shortly (see 7.6.2, the case of F_n^I).

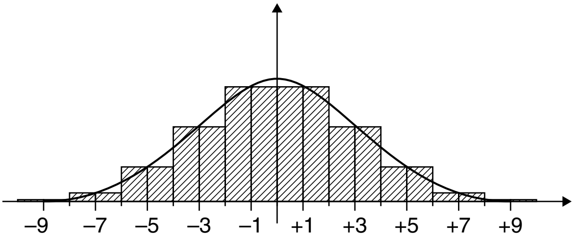

Figure 7.5 The binomial distribution: Heads and Tails $\left(p = \frac{1}{2}\right)$ with n tosses, $n = 9$. The possible values for the gain run through the ten odd numbers from −9 to +9, and the height of the column indicates the probability concentrated on each of these numbers. To give a more expressive picture, the values are assigned uniformly within ±1 of each point; this makes much clearer the approach of the binomial to the normal distribution, which will, in fact, be shown to be the limit distribution (as $n \to \infty$).

If we were, in fact, to consider the representation in terms of the natural scale (the gain Y_n, or the number of successes S_n) it would flatten out more and more, since the deviation behaves like $\sqrt{n}$. (One could think in terms of Bittering's apparatus, in which less and less sand would remain in each section as one continued to overturn it; see Figure 7.4: on the other hand, a large number of sections would be needed if one were to continue for very long.) In contrast to this, if one were to represent it in relative terms (the mean gain per toss, Y_n/n, or the frequency, S_n/n) the curve would shrink like $1/\sqrt{n}$, and would rise up to a peak in the centre. The remainder outside this central interval would tend to zero by virtue of the Tchebychev inequality. An appropriate choice is somewhere in between; as we have seen, one can take $Y_n/\sqrt{n}$ (i.e. the standardized deviation, both of the gain, and of the frequency from its prevision, $\frac{1}{2}$).

A somewhat more detailed study of the distribution of Heads and Tails will show us straightaway that the convergence to the normal distribution which is suggested by a visual inspection does, in fact, take place. In this case, too, however, the conclusions are valid more generally. They are valid not only for any binomial process with $p \neq 0$ (the effect of any asymmetry tends to vanish as n increases[32]) but also for sums of arbitrary, independent random quantities, provided that certain conditions (which will be given at the end of the chapter) are satisfied.

7.6.2. The limit distribution F of a sequence of distributions is to be understood in the sense defined in Chapter 6, 6.7.1: $F_n \to F$ means that, in terms of the distribution functions, $F_n(x) \to F(x)$ at all but at most a countable set of points (more precisely, at all but the possible discontinuity points of $F(x)$). This does not imply, of course, that if the densities exist we must necessarily also have $f_n(x) \to f(x)$; even less does it imply that if the densities are themselves differentiable we must have $f_n'(x) \to f'(x)$. Conversely, however, it is true that these properties do imply the convergence of the distributions (indeed, in a stronger and stronger, and intuitively meaningful way; one only has to think in terms of the graph of the density function).

[32] We mention this case explicitly since many people seem to doubt it (notwithstanding the fact that it is clearly covered by the general theorem). Perhaps this is the result of a misleading prejudice deriving from too much initial emphasis on Heads and Tails (?).

In our case, we can facilitate the argument by first reducing ourselves to the case just mentioned (albeit with a little trickery along the way). In fact, our distributions are discrete, standardized binomial with $p=\frac{1}{2}$ and hence with probabilities $p_h=\binom{n}{h}/2^n$ concentrated at the points $x_n=(2h-n)/\sqrt{n}$ (distance $2/\sqrt{n}$ apart, lying between $\pm\sqrt{n}$). In order to obtain a distribution admitting a density, it is necessary to distribute each mass, p_h, uniformly over the interval $x_h \pm 1/\sqrt{n}$; or, alternatively, with a triangular distribution on $x_h \pm 2/\sqrt{n}$. In this way, we obtain a continuous density (in the first case, the density is a step function; in the second, the derivative is a step function): we shall denote these two distributions by F_n^I and F_n^{II}, respectively.

We can also give a direct interpretation in terms of random quantities. The distributions arise if, instead of considering $Y_n/\sqrt{n}$, we consider $(Y_n + X)/\sqrt{n}$, where X is a random quantity independent of Y_n, having the appropriate distribution (in the cases we mentioned, $f(x)=\frac{1}{2}(|x|\leqslant 1)$, or $f(x)=\frac{1}{4}(2-|x|)$ $(|x|\leqslant 2)$, respectively[33]). We observe immediately that, since no mass is shifted by more than $1/\sqrt{n}$ (or $2/\sqrt{n}$, respectively) in one direction or the other, F_n^I and F_n^{II} will, for each x, from some n = N on, lie entirely between $F_n(x - \varepsilon)$ and $F_n(x + \varepsilon)$ (in fact, it suffices that $\varepsilon > 2/\sqrt{n}$; i.e. $n > \mathrm{N} = 4/\varepsilon^2$). It follows that, so far as the passage to the limit is concerned, it will make no difference if we use these variants in place of the actual F_n (for notational convenience, we shall not make any distinctions in what follows; we shall simply write F_n. The change in the standard deviation also makes no difference, and can be obtained immediately, without calculation, from the previous representation: X has standard deviation $1/\sqrt{3}$ in the case of a uniform distribution (between ± 1), and $\sqrt{(2/3)}$ for a triangular distribution (between ± 2), and it therefore follows that the addition of X either changes the standard deviation to $\sqrt{(1 + 1/3n)}$ or to $\sqrt{(1 + 2/3n)}$ (i.e., asymptotically to $1 + 1/6n$ and $1 + 1/3n$).

Having given these basic details, we can proceed rather more rapidly, arguing in terms of the more convenient modified distribution.

By distributing the mass p_h uniformly over the interval $x_h \pm 1/\sqrt{n}$, we obtain a density $f_n(x_h)=p_h/(2/\sqrt{n})=\frac{1}{2}p_h\sqrt{n}=\frac{1}{2}\sqrt{n}\binom{n}{h}/2^n$. By distributing it in a triangular fashion, the density at x_h remains the same, but, in every interval $[x_h, x_{h+1}]$, instead of preserving in the first and second half the values of the first and second end-points, respectively, it varies linearly (the graph = the jagged line joining the ordinates at the points $x = x_h$). In fact, the contribution of p_h decreases from x_h on, until it vanishes at x_{h+1} (and the contribution of p_{h+1} behaves in a symmetric fashion).

In the interval $x_h < x < x_{h+1}$, the derivative of the density, $f_n'(x)$, will therefore be constant:

$$f_n'(x)=(p_{h+1}-p_h).\frac{1}{2}\sqrt{n}/(2/\sqrt{n})=\frac{1}{4}n(p_{h+1}-p_h). \tag{7.21}$$

Recalling from 7.4.2 that $p_{h+1}/p_h = (n - h)/(h + 1)$, and from the expressions for $f_n(x_h)$ and x_h that $h=\frac{1}{2}(n+x_h\sqrt{n})$ (and similarly for x_{h+1}), we have the two alternative expressions

[33] In the first case, X has a uniform distribution over $|x| \leqslant 1$; in the second case, $X = X_1 + X_2$, where X_1 and X_2 are stochastically independent, and each has this uniform distribution.

$$f_n'(x) = p_h . \frac{n}{4}\left(\frac{n-h}{h+1} - 1\right) = p_h . \frac{1}{4} n \ . \frac{n-2h-1}{h+1}$$

$$= \frac{1}{2}\sqrt{n}\frac{-x_h\sqrt{n}-1}{\frac{1}{2}\left(n + x_h\sqrt{n}\right)+1} f_n(x_h) = -x_h . f_n(x_h) . \frac{1+1/\left(x_h\sqrt{n}\right)}{1+\left(x_h/\sqrt{n}\right)+(2/n)},$$

and similarly,

$$f_n'(x) = p_{h+1} . \frac{1}{4} n\left(1 - \frac{h+1}{n-h}\right) = -x_{h+1} \cdot f_n(x_{h+1}) \cdot \frac{1-1/\left(x_{h+1}\sqrt{n}\right)}{1-\left(x_{h+1}/\sqrt{n}\right)+(2/n)}.$$

This proves that the logarithmic derivative $f_n'(x)/f_n(x)$ (which clearly has its own extreme values to the right of the left-hand end-point, and to the left of the right-hand end-point) always satisfies the equation

$$f_n'(x)/f_n(x) = \frac{\mathrm{d}}{\mathrm{d}x}\log f_n(x) = -x\left[1+\varepsilon(x)\right] \tag{7.22}$$

(where, as n increases, $\varepsilon(x)$ tends uniformly to 0 in any finite interval which has a neighbourhood of the origin removed; e.g. we have $\varepsilon(x) < \varepsilon$, for some n, throughout the interval $2\varepsilon/\sqrt{n} < |x| < \sqrt{n}/2\varepsilon$; the apparent irregularity at the origin merely stems, however, from the fact that both x and $f'(x)$ go to zero, and the equation is automatically satisfied without there being any need to consider the ratio).

The limit distribution must, therefore, satisfy

$$f'(x)/f(x) = -x, \tag{7.23}$$

from which we obtain

$$\log f(x) = -\frac{1}{2}x^2 + \text{const.}, \quad f(x) = K\mathrm{e}^{-\frac{1}{2}x^2} \quad \left(K = 1/\sqrt{(2\pi)}\right). \tag{7.24}$$

The conclusion is therefore as follows: *the standardized binomial distribution* (the case of Heads and Tails, $p = \frac{1}{2}$) *tends, as n → ∞, to the standardized normal distribution.* The same conclusion holds, however, in more general cases and, because of its importance, is known as the *central limit theorem* of the calculus of probability. We see immediately that the conclusion holds in the binomial case for $p \neq \frac{1}{2}$ (except that we now require different coefficients in order to obtain the *standardized* form).

7.6.3. It is convenient at the beginning to dwell upon the rather special example of Heads and Tails, since this provides an intuitive and straightforward illustration of many concepts and techniques, which themselves have a much broader compass, but whose essential meaning could otherwise be obscured by the technical details of the general case.

The proof we have just given (based on a technique used by Karl Pearson for this and other examples) is probably the easiest (even more so if one omits the details of the inequalities and simply makes the heuristic observation that, for large enough n,$f'(x)/f(x)$ is practically equal to $-x$).

Remark. Geometrically, this means that the *subtangent* – $1/x$ of the graph of $y = f(x)$ is inversely proportional to the abscissa. The *tail* beyond x is, approximately, an exponential distribution, with density $f(\xi) = K\,\mathrm{e}^{-x\xi}$[34] and prevision $1/x$; this is, in fact, as we can see asymptotically from equation 7.20), the prevision of the excess of X over x (provided it does exceed it). This means, essentially, that if an error X (with a standardized normal distribution) exceeds some large given value x, it is almost certain that it exceeds it by very little (about $1/x$): for example, if it exceeds 10σ (or 100σ), we can expect that it exceeds it by $\sigma/10$ (or $\sigma/100$).

Note that this is precisely what happens for the deviations of Heads and Tails (see the footnote to equation 7.4 of Section 7.4.2), provided we make appropriate allowances for the discreteness. If we know that Heads have occurred in more than 75% of the trials, the probabilities that it has occurred 1, 2, 3, 4, or more than 4 times beyond this limit are 0·67, 0·22, 0·074, 0·025, 0·012, respectively, no matter how many tosses n have been made. This means that for $n = 100$ it is almost certain that (with the probabilities given above) the number of successes is one of 76, 77, 78, 79, whereas, for $n = 1000$, the same holds for 751, 752, 753, 754, for $n = 1{,}000{,}000$, for 750,001, 750,002, 750,003, 750,004!

We shall present other (and more general) proofs of this theorem later and it will be instructive to see it tackled from different standpoints. For the moment, however, we shall consider a useful corollary of it.

Using the fact that $f_n(x) \simeq f(x)$, and recalling the relation with p_h, we find that

$$\omega_h^{(n)} = p_h \simeq \left(2/\sqrt{n}\right) f_n(x_h) \simeq \left(2/\sqrt{n}\right) f(x_h) = \sqrt{(2/\pi n)}\exp\left[-\frac{1}{2n}(2h-n)^2\right]. \qquad (7.25)$$

In particular, for $x = 0$, we obtain the maximum term, that is the central one ($h = \frac{1}{2}n$ if n = even, or either of $h = \frac{1}{2}(n \pm 1)$ if n = odd). We shall always denote this by a special symbol, u_n, and the formula we have arrived at gives the asymptotic expression $u_n \simeq \sqrt{(2/\pi n)}$; that is, in figures, $u_n \simeq 0{\cdot}8/\sqrt{n}$ (this makes clear the meaning of the coefficient $\sqrt{(2/\pi)}$, which it is important to keep in mind). In fact, the probability u_n (the maximum probability among the $\omega_h^{(n)}$ of the Heads and Tails case) will crop up in many problems (a partial summary of which will be given in Chapter 8, 8.7.4). For the present, we shall just indicate a few of its properties.

In fact, we have

$$\begin{aligned} u_n = u_{2m} &= \mathbf{P}(Y_n = 0) = \omega_m^{(n)} && \text{for } n = 2m = \text{even},\\ u_n = u_{2m-1} &= \mathbf{P}(Y_n = 1) = \mathbf{P}(Y_n = -1) && \\ = \omega_m^{(n)} &= \omega_{m-1}^{(n)} = u_{2m} && \text{for } n = 2m-1 = \text{odd}. \end{aligned} \qquad (7.26)$$

The equality of the u_n for successive pairs of values (each odd one with the next even one) is obvious from the definition. In order that the gain after $2m$ tosses be zero, it is necessary that it was either +1 or −1 at the preceding toss and that the final toss had

[34] We have exp $\left\{-\frac{1}{2}(x-\xi)^2\right\} = \exp(-\frac{1}{2}x^2)\exp(-x\xi)\exp(-\frac{1}{2}\xi^2)$, but only the first factor remains because the second is constant (with respect to ξ), and is incorporated into K, and the third is $\simeq 1$ (for small ξ).

the outcome required to bring it to 0; both possibilities have probability $u_{2m-1}\cdot\frac{1}{2}$, and their sum gives

$$u_{2m} = 2\left(\frac{1}{2}u_{2m-1}\right) = u_{2m-1}.$$

The same argument can be carried out for the binomial coefficients by applying Stiefel's formula. The central term, $\binom{2m}{m}$, for n = $2m$ = even, is the sum of the two adjacent ones which are themselves equal,

$$\binom{2m-1}{m-1} = \binom{2m-1}{m},$$

and is therefore twice their value; in order to obtain the probability, however, we must divide by 2^{2m} rather than by 2^{2m-1}, so u_{2m} = u_{2m-1}. We obtain, therefore,

$$u_{2m-1} - u_{2m} = \binom{2m}{m}/2^{2m} = \frac{(2m)!}{2^{2m}(m!)^2} \simeq \surd(2/\pi n) \simeq 0\cdot 8/\surd n. \tag{7.27}$$

7.6.4. We see from this that the factor √(2/π;), hitherto regarded simply as the normalization factor for the standardized normal distribution, also has a link with the combinatorial calculus. This connection is given by Stirling's formula, which provides an asymptotic expression for the factorial and which enables us to arrive at the central limit theorem for the binomial distribution by a different route (one that is more laborious but is often used and is, in any case, useful to know).

Stirling's formula expresses n! as follows:

$$n! = n^n \mathrm{e}^{-n}\surd(2\pi n)(1+\varepsilon_n)\left(\text{where } \varepsilon_n \to 0; \text{more precisely}, 0 < \varepsilon_n < 1/1\ \ 1n\dagger\right)^{35}. \tag{7.28}$$

Since the formula is used so often, we shall give a quick proof of it. We have

$$\begin{aligned}\log n! &= \log 2 + \log 3 + \ldots + \log n \\ &\simeq \int_{\frac{3}{2}}^{n+\frac{1}{2}} \log x\,\mathrm{d}x = \left[x\log x - x\right]_{\frac{3}{2}}^{n+\frac{1}{2}} \\ &= \left(n+\frac{1}{2}\right)\log n - n + \text{const.},\end{aligned} \tag{7.29}$$

and we observe that the difference between the sum and the integral converges (substituting log n in place of

$$\int_{n-\frac{1}{2}}^{n+\frac{1}{2}} \log x\ \ \mathrm{d}x,$$

[35] Note that, if we neglect ε_n, Stirling's formula gives n! with smaller and smaller *relative* error, but with greater and greater *absolute* error (i.e. the *ratio* tends to 1, but the *difference* between n! and the approximation tends to +∞). In practical terms, for $n \simeq 10^k$ we have n! with about the first k + 1 digits correct; but n! (for large k) has about 10^k digits, and the error has not many less. In any case, what matters in applications is the relative approximation, and this is adequate even for small values.

we see immediately that we have an error of order $1/n^2$). From this, it follows that $n! \simeq Kn^{n+\frac{1}{2}}\,e^{-n}$ (a result known to De Moivre). As for the fact that $K = \sqrt{(2\pi)}$ (discovered by Stirling in 1730), we shall consider it as being established heuristically by virtue of the fact that, were we to leave it indeterminate, the limit of the $f_n(x)$ would be given by

$$f(x) = (1/K)e^{-\frac{1}{2}x^2}$$

and we know that this multiplicative factor must be $1/\sqrt{(2\pi)}$.

Let us just evaluate u_n by this method (n even: $n = 2m$): we obtain

$$\begin{aligned} u_n &= \binom{2m}{m}/2^{2m} = \frac{(2m)!}{2^{2m}(m!)^2} \simeq \frac{2^{2m}m^{2m}e^{-2m}\sqrt{(2\pi 2m)}}{2^{2m}\left[m^m e^{-m}\sqrt{(2\pi m)}\right]^2} \\ &= \frac{1}{\sqrt{(\pi m)}} = \sqrt{(2/\pi n)} = 0\cdot 8\sqrt{n}. \end{aligned}$$

In order to evaluate

$$\omega_{m+k}^{(n)} = \frac{(2m)!}{2^{2m}(m-k)!(m+k)!} = \omega_m^{(n)}\frac{(m!)^2}{(m-k)!(m+k)!},$$

it is more convenient to make use of an alternative form of Stirling's formula, one which will turn out to be useful in a number of other cases. It is based on evaluating products of the form $(1 + a)$ $(1 + 2\mathrm{a})\ldots$ $(1 + ka)$, with k large, and $ka = c$ small; in our case, $[m!/(m - k)!]/[(m + k)!/m!]$ can be written as

$$\left[1.(1-a)(1-2a)\ldots(1-(k-1)a)\right]\Big/\left[(1+a)(1+2a)\ldots(1+ka)\right]$$

by dividing both ratios by m^k, and setting $1/m = a$.

Taking the logarithm, we have

$$\begin{aligned} \log\prod_{h=1}^{k}(1+ah) &= \sum_{h=1}^{k}\log(1+ah) \simeq \frac{1}{a}\int_1^{\lambda}\log x\,dx \\ &= \frac{1}{a}\left[(1+\lambda)\log(1+\lambda)-\lambda\right], \end{aligned}$$

with $\lambda = (k+\frac{1}{2})a$,[36] and, expanding in a series, we have

$$\begin{aligned} \log\prod_{h=1}^{k}(1+ah) &= \frac{\lambda^2}{2a}\left(1-\frac{\lambda}{3}+\frac{\lambda^2}{6}-\frac{\lambda^3}{10}+\ldots\pm\frac{\lambda_n}{\frac{1}{2}(n+1)(n+2)}\mp\ldots\right) \\ &\simeq \frac{1}{2}\left(k+\frac{1}{2}\right)^2 a. \end{aligned}$$

[36] The simpler form $\lambda = ka$ is practically equivalent to the effect of an individual evaluation. In the case of products or ratios of a number of expressions of this kind, however, it can happen (and does in the example of Section 7.4.3) that it is the contributions deriving from the '+½' which are important, because the main contributions cancel out.

It follows that

$$\left(1+a\right)\left(1+2a\right)\ldots\left(1+ka\right) \simeq \mathrm{e}^{\frac{1}{2}\left(k+\frac{1}{2}\right)^2 a} \simeq \mathrm{e}^{\frac{1}{2}ak^2}. \tag{7.30}$$

In our case, with $a = \pm\, 1/m$, the two products equal $\mathrm{e}^{\pm\frac{1}{2}ak^2} = \mathrm{e}^{\pm k^2/2m}$ and their ratio is

$$\mathrm{e}^{-k^2/2m} / \mathrm{e}^{k^2/2m} = \mathrm{e}^{-k^2/m} = \mathrm{e}^{-\left(h-m\right)^2/m} = \mathrm{e}^{-\left(2h-n\right)^2/2n}$$

(since $k = h - m$ and $m = \frac{1}{2}n$). We thus obtain the result (which, of course, we already knew).

7.6.5. *Relation to the diffusion problem.* We give here a suggestive argument (due to Pólya), which is entirely heuristic, but is very useful as a basis for discussions and developments. The relation between random processes of the kind we have just exemplified with Heads and Tails and diffusion processes, which we shall meet later, will, in fact, provide a basis for interpreting the latter and even identifying them with the kinds of process already studied. The Wiener–Lévy process (see Chapter 8) can, with reference to our previous work, be thought of as a Heads and Tails process involving an enormous number of tosses with very small stakes, taking place at very small time intervals. This process has also been referred to (by P. Lévy) as the *Brownian motion* process, because it can be used (although only for certain aspects of the problem) to represent and study the phenomenon of the same name (which is, as is well known, a diffusion process).

The Heads and Tails process can be thought of as a diffusion process in which a mass (a unit mass, initially – i.e. at $t = 0$ – concentrated at the origin) moves, with respect to time t, through the lattice of Figure 7.2, splitting in half at each intersection (encountered at times t = integer). The mass (which represents the probability) would, according to this representation, divide up in a certain (i.e. deterministic) manner, and, formally, everything goes through (indeed, it will be even simpler than this).

A more meaningful interpretation, however, and one more suited to our purpose, derives from consideration of a random process of the statistical type. Assume that, initially, a very large number of particles (N, say) are concentrated at the origin, and move at equal and constant rates to the right, through the lattice. At each time instant t = integer, they meet an intersection, and each chooses its direction independently of the others. Equivalently, we could think of them as moving with constant speed on the y-axis, choosing directions at random at each time instant t = integer (i.e. each time a point y = integer is reached); alternatively, we could think of them at rest, but making a jump of ±1 at each t = integer.

Taking the total mass = 1, the mass crossing a given point can no longer be determined with certainty: where, in the deterministic case, it was ω, we can now only say that we have prevision ω and that the number of particles has prevision $N\omega$, but could take any value h, lying between 0 and N, with probability $\binom{N}{h}\omega^h(1-\omega)^{N-h}$. If we want to give a rough idea of what happens, we could say (quoting the prevision ± the standard deviation) that the number of particles will be

$$N\omega \pm \surd\left[N\omega\left(1-\omega\right)\right]$$

($\simeq N\omega \pm \surd(N\omega)$ for small ω; the Poisson approximation).

This is what we are interested in: a normal distribution being attained as a result of a statistical diffusion process.

For the purposes of the mathematical treatment (whatever the interpretation), the mass crossing the point (vertex) (t, y), where t and y are integer, both even or both odd, is, as usual,

$$\mathbf{P}(Y_t = y) = \omega_{(t+y)/2}^{(t)} = \omega(t,y),$$

given by one half of that which has crossed $(t - 1, y - 1)$ or $(t - 1, y + 1)$:

$$\omega(t,\ y) = \frac{1}{2}\Big[\omega(t-1,\ y-1) + \omega(t-1,\ y+1)\Big].$$

The notation $\omega(t, y)$ has been introduced in order to allow us to think of the function as defined everywhere (no matter what the interpretation), even on those points where it has no meaning in the actual problem; in particular, for t and y integer, but $t + y$ odd, like $\omega(t - 1, y)$. Subtracting this value from both sides of the previous equation, we obtain

$$\Delta_t \omega = \frac{1}{2}\Delta_y^2 \omega \tag{7.31}$$

and, in the limit,

$$\frac{\partial \omega}{\partial t} = \frac{1}{2}\frac{\partial^2 \omega}{\partial y^2}, \tag{7.32}$$

provided that (taking the units of t and y to be very small) one considers it legitimate to pass from the discrete to the continuous.

Let us restrict ourselves here to simply pointing out that, in this way, one arrives at the correct solution. In fact, the general solution of the *heat equation*, (7.32), is given by

$$\omega(t,y) = (K/\sqrt{t})\mathrm{e}^{-\frac{1}{2}y^2/t}, \tag{7.33}$$

a well-known result that can easily be verified.

7.6.6. The form of the normal distribution is well known, and is given in Figure 7.6 (where we show the density $y = f(x)$). We also provide a table of numerical values for both the density and the distribution function (the latter giving the probabilities of belonging to particular half-lines or intervals).

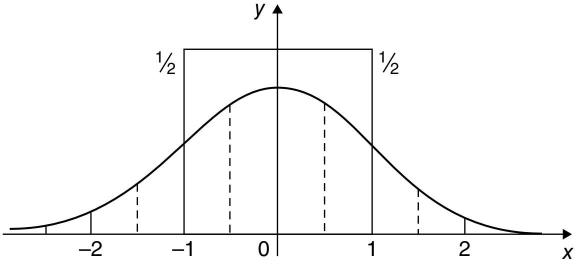

Figure 7.6 The standardized normal distribution ($m = 0, \sigma = 1$): the density function. The subdivisions (0, ±1, ±2, ±3) correspond to σ, 2σ, 3σ; at ±1 we have points of inflection, between which the density is convex. The rectangle of height $\frac{1}{2}$ shows, for comparative purposes, the uniform distribution on the interval [−1, +1] (which might well be called the 'body' of the distribution; see Chapter 10, 10.2.4). Note that the vertical scale is, in fact, four times the horizontal one, in order to avoid the graph appearing very flat (as it is, in fact), and hence not displaying the behaviour very clearly.

We shall confine ourselves to calling attention to a few points of particular importance.

The density function is symmetric about its maximum, which is at the origin, and decreases away from it, being convex (upwards) in the interval [−1, +1], and concave outside this interval. As $x \pm \infty$, it approaches the x-axis, the approach being very rapid, as we already pointed out in 7.6.3 (the 'tails' are 'very thin'). In fact, the subtangent decreases, in our case, like l/$|x|$ (if $f(x)$ tends to 0 like a power, $|x|^{-n}$, with n arbitrarily large, the subtangent, in absolute value, increases indefinitely as $|x|$ increases; if the function decreases exponentially, the subtangent is constant).

The graph has two points of inflection at ±1 (corresponding to the change from convexity to concavity that we already mentioned). The ordinate at these points is about 0·6 of the maximum value (the table gives 0·60652, but it is better to keep an approximate round figure in mind; this is enough to prevent one from making the all too usual distortions when sketching it – on the blackboard, for example). The subtangents are equal to ∓1; that is the slope is such that the tangents cross the x-axis at the points ±2.

Since the tails are 'very thin', it is clear that the probabilities of the occurrence of extreme values beyond some given x are, in the case of the normal distribution, much smaller than is usual (in the case of densities decreasing like powers, or exponentially). They are, therefore, much smaller than the values provided by Tchebychev's inequality, which is valid under very general conditions.

We give below a few examples of the probabilities of $|X|$ exceeding $k\sigma$ (or, in the standardized case, $\sigma = 1$, of exceeding k), for $k = 1, 2, 3$ and $3\frac{1}{2}$:

Absolute value greater than:		σ	2σ	3σ	$3\frac{1}{2}\sigma$
Probability	normal distribution:	31·74%	4·55%	0·27%	0·05%
	Tchebychev inequality:	⩽100·00%	⩽25·00%	⩽11·11%	⩽8·16%

The table is not only useful for numerical applications but it should also be used in order to commit to memory a few of the significant points (e.g. a few ordinates and, more importantly, the areas corresponding to abscissae 1, 2 and 3; i.e. to σ, 2σ and 3σ).

The reader is invited to refer to equation 7.20 in Section 7.5.4, and to the *Remarks* of Section 7.6.3, where we looked at asymptotic expressions for such probabilities ($\simeq K\mathrm{e}^{-\frac{1}{2}x^2}/x$, $K = 1/\sqrt{(2\pi)} \simeq 0{\cdot}40$[37]), and at the order of magnitude for possible exceedances (prevision $\simeq 1/x$).

Table of values for the standardized normal (Gaussian) distribution

Abscissa	Ordinate (density)		Area ($\int f(x)\,dx$) in %		
X	$f(x) = \dfrac{1}{\sqrt{(2\pi)}}\mathrm{e}^{-\frac{1}{2}x^2}$	$f(x)$ as % of central ordinate	from x to $+\infty$	in the individual intervals given	2 × (5)
(1)	(2)	(3)	(4)	(5)	(6)
0·0	0·398942	100·0	50·0	19·15	38·30
0·1	0·396952	99·50	46·0172		
0·2	0·391043	98·02	42·0740		
0·3	0·381388	95·60	38·2089		
0·4	0·368270	92·31	34·4978		
0·5	0·352065	88·25	30·8538		

[37] Writing $K(1 + \Theta/x^2)$, with $0 \leqslant \Theta \leqslant 1$, in place of K, we have an exact bound (and $\Theta \sim 1 - 3/x^2$).

Abscissa	Ordinate (density)		Area $(\int f(x)\,dx)$ in %		
X	$f(x)=\frac{1}{\sqrt{(2\pi)}}e^{-\frac{1}{2}x^2}$	$f(x)$ as % of central ordinate	from x to $+\infty$	in the individual intervals given	2 × (5)
(1)	(2)	(3)	(4)	(5)	(6)
0·6	0·333225	83·53	27·4253	14·98	29·96
0·7	0·312254	78·27	24·1964		
0·8	0·289692	72·61	21·1855		
0·9	0·266085	66·70	18·4060		
1·0	0·241971	60·652	15·8654		
1·1	0·217852	54·61	13·5666	9·19	18·38
1·2	0·194186	48·68	11·5070		
1·3	0·171369	42·96	9·6800		
1·4	0·149727	37·53	8·0756		
1·5	0·129518	32·47	6·6807		
1·6	0·110921	27·80	5·4799	4·405	8·810
1·7	0·094049	23·57	4·4565		
1·8	0·078950	19·79	3·5930		
1·9	0·065616	16·45	2·8717		
2·0	0·053991	13·53	2·2750		
2·5	0·017528	4·39	0·6210	2·140	4·280
3·0	0·004432	1·11	0·1350		
3·5	0·0008727	0·22	0·02326	0·135	0·270
∞	0·0	0	0·0		

Since the binomial distribution (and many others) approximates, under given conditions, as n increases to the normal, the table of the latter can also be used in other contexts (with due care and attention[38]). Such tables are often used in the case of empirical distributions (statistical distributions), under the confident assumption that the latter behave (at least approximately) like normal distributions. It is easily shown (see Section 7.6.9) that this confidence is often not, in fact, justified.

7.6.7. In order to deal with certain other instructive and important features of the normal distribution, we shall have to refer to the multidimensional case (either just two dimensions, the plane, or some arbitrary number n; or even the asymptotic case, $n \to \infty$).

It will suffice to limit our discussion to the case of spherical symmetry, where the density has the form

$$f\left(x_1,x_2,\ldots,x_r\right)=K\exp\left(-\frac{1}{2}\rho^2\right),\quad \rho^2=x_1^2+x_2^2+\ldots+x_r^2.$$

This corresponds to assuming the X_h to be standardized ($m=0$, $\sigma=1$) and stochastically independent (for which, in the case of normality, it is sufficient that they be uncorrelated). In fact, we can always reduce the general situation to this special case provided we apply to S_r the affine transformation that turns the covariance ellipsoid into a 'sphere' (see Chapter 4, 4.17.6). In other words, we change from the X_h to a set of Y_k which are

[38] If one were not careful, one might conclude that the probability of obtaining more than n Heads in n tosses(!) is very small, but not zero (about $2{\cdot}4\times10^{-23}$ for $n=100$; about 10^{-2173} for $n=10{,}000$).

standardized and uncorrelated (and are linear combinations of the X_h). We shall have more to say about this later (Chapter 10, 10.2.4).

For the moment, let us evaluate the normalizing constant of the standardized normal distribution (which we have already stated to be $K = 1/\surd(2\pi)$). Integrating over the plane, we obtain

$$\iint \mathrm{e}^{-\frac{1}{2}x^2}\mathrm{e}^{-\frac{1}{2}y^2}\,\mathrm{d}x\,\mathrm{d}y = \int \mathrm{e}^{-\frac{1}{2}\rho^2}\rho\,\mathrm{d}\rho \int \mathrm{d}\theta = 2\pi,$$

which is also equal to $\left[\int \mathrm{e}^{-\frac{1}{2}x^2}\mathrm{d}x\right]^2$. It follows that $K = 1/2\pi$ in the plane ($r = 2$), $K = (2\pi)^{-\frac{1}{2}}$ over the real line ($r = 1$), and, for general r, $K = (2\pi)^{-\frac{1}{2}r}$.[39]

There are another two important, interesting properties to note. They involve the examination of two conditions, closely linked with one another, each of which provides a meaningful characterization of the normal distribution. In both cases, it is sufficient to deal with the case of the plane.

The first of them is summarized in the following: the only distribution over the plane which has circular symmetry, and for which the abscissa X and the ordinate Y are stochastically independent (orthogonal), is that in which X and Y have normal distributions with equal variances (and zero prevision, assuming the symmetry to be about the origin). The second property concerns the *stability* of distributions (which we discussed in Chapter 6, 6.11.3). If we require, in addition, that the variance be finite, then stability is the *exclusive* property of the normal distribution.

For the first property, if we denote by $f(\cdot,\cdot), f_1(\cdot), f_2(\cdot)$ the joint density for (X, Y), and the marginal densities for X and Y, respectively, the given conditions may be expressed as follows:

a) *rotational symmetry;* $f(x, y)$ = const, for $x^2 + y^2 = \rho^2$ = const., from which it follows immediately that $f(x, y) = f(\rho, 0) = f(0, \rho)$ for $\rho = \surd(x^2 + y^2)$;
b) independence; $f(x, y) = f_1(x)f_2(y)$.

In our case, given the symmetry, we can simply write $f(\cdot)$ instead of $f_1(\cdot)$ and $f_2(\cdot)$, and hence obtain a single condition,

$$f(x,y) = f(x)f(y) = f(0)f(\rho)$$

In other words,

$$\frac{f(x)}{f(0)}\frac{f(y)}{f(0)} = \frac{f(\rho)}{f(0)},$$

and, if we put $f(x)/f(0) = \psi(x^2)$, this gives the functional equation

$$\psi(x^2)\psi(y^2) = \psi(\rho^2) = \psi(x^2 + y^2).$$

[39] We should make it clear (because it is customary to do so – it is, in fact, obvious) that the integral of a positive function taken over the plane does not depend on how one arrives at the limit (by means of circles, $\rho < R$, or squares, $|x| \vee |y| < R$, or whatever); it is always the supremum of the values given on the bounded sets.

Taking logarithms, this gives the additive form

$$\log\psi\left(x^2+y^2\right)=\log\psi\left(x^2\right)+\log\psi\left(y^2\right).$$

Under very weak conditions, which are usually satisfied, this implies linearity (e.g. it is sufficient that ψ is non-negative in the neighbourhood of some point; here, this holds over the whole positive real line). It follows that

$$\log\psi\left(x^2\right)=kx^2,\quad \psi\left(x^2\right)=\mathrm{e}^{kx^2},\quad f(x)=f(0)\mathrm{e}^{kx^2}$$

(and, normalizing, that $k= -l/2\sigma^2$ and $f(0)= 1/\sqrt{(2\pi)}\sigma)$; the required property is therefore established.

In certain cases, this property is in itself sufficient to make the assumption of a normal distribution plausible. A celebrated example is that of the distribution of the velocity of the particles in Maxwell's kinetic theory of gases. If one assumes: (a) *isotropy* (the same distribution for components in all directions) and (b) stochastic *independence* of the orthogonal components, then the distribution of each component is normal (with zero previsions and equal variances). In other words, the distribution of the velocity vector is normal and has spherical symmetry (with density $K\mathrm{e}^{-\frac{1}{2}\rho^2/\sigma^2}$).

Given the assumptions, the above constitutes a mathematical proof. But however necessary they are as a starting point, the question of whether or not these (or other) assumptions should be taken for granted, or regarded as more or less plausible, is one which depends in part upon the actual physics, and in part upon the psychology of the author concerned.

The second property referred to above reduces to the first one. We must first of all restrict ourselves to the finite variance case (otherwise, we already know the statement to be false; see the stable, Cauchy distribution, mentioned at the end of Chapter 6, 6.11.3), and we might as well assume unit variance. We therefore let $f(x)$ denote the density of such a distribution (with $m = 0$, $\sigma = 1$), and X and Y be two stochastically independent random quantities having this distribution.

In order for there to be stability, $Z = aX + bY$ must, by definition, have the same distribution (up to a change of scale, since $\sigma^2 = a^2 + b^2$). If, by taking $a^2 + b^2 = 1$, we make $Z = X\cos\alpha + Y\sin\alpha$, we can avoid even the change of scale, and we can conclude that all projections of the planar distribution, $f(x, y) = f(x)f(y)$, in whatever direction, must be the same. In other words, the projections must possess circular symmetry and it can be shown that a necessary condition for this (and clearly a sufficient one also) is that the density has circular symmetry (as considered for the first property).[40]

The conclusion is, therefore, the same: the property characterizes the normal distribution. The result contains within it an implicit justification (or, to be more accurate, a partial justification) of the 'central limit theorem'. In fact, if the distribution (standardized, with $\sigma = 1$) of the gain from a large number of trials at Heads and Tails follows, in practice, some given distribution, then the latter must be stable (and the same is true for any other case of stochastically independent gains). It is sufficient to note that if Y' and Y'' are the gains from large numbers of trials, n' and n'', respectively, then, *a fortiori*,

[40] This is intuitively entirely 'obvious'. The proof, which is rather messy if one proceeds directly, follows immediately from the properties of characteristic functions of two variables (Chapter 10, 10.1.2).

$Y = Y' + Y''$ is the gain from $n = n' + n''$ trials. If the two independent summands belong to the limit distribution family, then so does their sum: this implies stability.

The justification is only partial because the above argument does not enable us to say whether, and in which cases (not even for that of Heads and Tails), there is convergence to a limit distribution. It does enable us to say, however, that *if a limit distribution exists* (with a finite standard deviation, and if the process is additive with independent summands – all obvious conditions) *it is necessarily the normal distribution.*

7.6.8. *An interpretation in terms of hyperspaces.* It is instructive to bear in mind, as an heuristic, but meaningful, interpretation, that which can be given in terms of hyperspaces. Compared with the previous example, it constitutes even less of a 'partial justification' of the appearance of the normal distribution under the conditions of the central limit theorem (sums of independent random quantities), but, on the other hand, it reveals how the result is often the same, even under very different conditions.

Let us begin by considering the uniform distribution inside the sphere (hypersphere) of unit radius in S_r, and the projection of this distribution onto the diameter, $-1 \leqslant x \leqslant +1$. The section at x has radius $\surd(l - x^2)$, 'area' equal to $[\surd(l - x^2)]^{r-1}$, and hence the density is given by

$$f(x) = K\left(1 - x^2\right)^{(r-1)/2}. \tag{7.34}$$

In particular, we have $K\surd(1 - x^2)$ for $r = 2$ (projection of the area of the circle); $K(1 - x^2)$ for $r = 3$ (projection of the volume of the sphere); and so on.

As r increases, the distribution concentrates around the origin (as happened in the case of frequencies at Heads and Tails). In order to avoid this and to see what, asymptotically, happens to the 'shape' of the distribution, it is necessary (again, as in the case of Heads and Tails) to expand it in the ratio 1: $\surd r$ (i.e. by replacing x by $x/\surd r$). We then obtain

$$f(x) = K\left(1 - \frac{x^2}{r}\right)^{(r-1)/2} \to K\mathrm{e}^{-\frac{1}{2}x^2}.$$

In the limit, this gives the normal distribution, but without any of the assumptions of the central limit theorem. What is more surprising, however, is that the same conclusion holds under circumstances even less similar to the usual ones. For example, it also holds if one considers a hollow sphere, consisting of just a small layer between $1 - \varepsilon$ and 1 ($\varepsilon > 0$ arbitrary). It is sufficient to note that the mass inside the smaller sphere contains $(1 - \varepsilon)^r$ of the total mass. This tends to 0 as r increases, and hence its contribution to the determination of the shape of $f(x)$ becomes negligible.

Well: the central limit theorem, also, can be seen as a special case of, *so to speak*, this kind of tendency for distributions in higher dimensions to have normally distributed projections onto a certain straight line.

The case of Heads and Tails shows that one obtains this projection (asymptotically, for large n) by projecting (onto the diagonal) a distribution of equal masses $\left(\frac{1}{2}\right)^n$ on the 2^n vertices of an n-dimensional hypercube. The same holds, however, for projections onto any other axis (provided it does not belong entirely to a face having only a small number of dimensions compared with n) and also if one thinks of the cube as a solid (with uniformly distributed mass inside it), or with a uniform distribution on the surface, and so on. To summarize; the interpretation in terms of hyperspaces holds in all cases where the central limit theorem holds (although it cannot be of any help in picking out these cases, except in those cases where there exists an heuristic argument by analogy with some known case).

More specifically useful is the conclusion that can be drawn in the opposite sense: namely, that of convergence to the normal distribution in many more cases than those, already numerous, which fall within the ambit of sums of independent random quantities, the case we are now considering (having started with the case of Heads and Tails). The solid cube does fall within such an interpretation (summands chosen independently and with a uniform distribution between ±1), but that of a distribution on the surface (or on edges, or m-dimensional faces) does not, and this would be even less true for the case of the hypersphere (solid, or hollow).

A wide-ranging generalization of the central limit theorem was given by R. von Mises,[41] and shows that even the distributions of nonlinear 'statistical functions' may be normal (under conditions which, in practice, are not very restrictive). Examples of nonlinear statistical functions of the observed values $X_1, X_2, \ldots, X_n$ are the means (other than the arithmetic mean, or their deviations from it), the moments and the functions of the moments (e.g. μ_3^2/μ_2^3, or $(\mu_4/\mu_2^2)-3$, where the μ_h are the moments about the mean and the expressions are used as indices of asymmetry and 'kurtosis', respectively; see Chapter 6, 6.6.6), the concentration coefficient (of Gini; see Chapter 6, the end of 6.6.3), and so on. In general, they are the functions that can be interpreted as functionals of the $F_n(x) = (1/n)\,\Sigma(X_h \leqslant x)$ (jump $1/n$ for $x = X_1, X_2, \ldots, X_n$), that is of the statistical distribution, under conditions similar to differentiability (i.e. of local linearity'). In essence (the actual formulation is quite complicated and involves long preliminary explanations before one can even set up the notation), one requires that the first derivative (in the sense of Volterra, for 'fonctions de ligne') satisfies the conditions for the validity of the 'central limit theorem' in the linear case and that the second derivative satisfies a complementary condition.

This generalization, wide ranging though it is, does not, however, include the cases that we considered in the hyperspace context. This emphasizes even further just how general is the 'tendency' for the normal distribution to pop up in any situation involving 'chaos'.

7.6.9. *Order out of chaos.* We shall postpone what we consider to be a *valid* proof of the central limit theorem until the next section (from a mathematical viewpoint it is a stronger result). Let us consider first the notion of 'order generated out of chaos', which has often been put forward in connection with the normal distribution (as well as in many other cases).

A general observation, which is appropriate at this juncture, concerns a phenomenon that often occurs in the calculus of probability; that of obtaining conclusions which are extremely precise and stable, in the sense that they hold unchanged even when we start from very different opinions or situations. This is the very opposite of what happens in other fields of mathematics and its applications, where errors pile up and have a cumulative effect, with the risk of the results becoming completely invalidated, no matter

[41] R. von Mises, *Selected Papers*, Vol. II, Providence (1964); see various papers, among which (pp. 388–394) the lectures given in Rome (Institute of Advanced Mathematics) provide one of the most up to date expositions (for an exposition of a more illustrative kind, see pp. 246–270).

how carefully the initial data were evaluated and the calculations carried out. This particular phenomenon compensates for the disadvantages inherent in the calculus of probability due to the subjective and often vague nature of the initial data. It is because of this peculiarity (which is, in a certain sense, something of a miracle, but which, after due consideration, can be seen, in a certain sense, as natural) that a number of conclusions appear acceptable to everyone, irrespective of inevitable differences in initial evaluations and opinions. This is a positive virtue, notwithstanding the drawbacks which stem from a too indiscriminate interpretation of it, leading one to accept as objective those things whose roots are, in fact, subjective, but have not been explicitly recognized as such.

In this connection, we shall now put forward an example that is basically trivial but, nonetheless, instructive (because it is clear what is going on; the central limit theorem is less self-evident). We return to case (*E*) of Section 7.2.2 and consider the probability of an odd number of successes out of n events, E_1, E_2,..., E_n. If we assume them to be stochastically independent with probabilities p_1, p_2,..., $p_{n,}$ the probability in question is given by

$$q_n = \frac{1}{2} - \prod_{h=1}^{n} (1 - 2p_h)$$

(which can be verified by induction). As n increases, the difference between q_n and $\frac{1}{2}$ decreases (in absolute value). In other words, if one is interested in obtaining a probability close to $\frac{1}{2}$, it is always better to add in (stochastically independent) events, whatever the probabilities p_h might be, because the above-mentioned difference is multiplied by $2(\frac{1}{2} - p_h)$, which is $\leqslant 1$ in absolute value, and the smaller the difference is, the closer p_h is to $\frac{1}{2}$: if $p_h = \frac{1}{2}$, the difference becomes zero (as we remarked at the time). Suppose we now consider a cube or a parallelepiped that we wish to divide into two equal parts as accurately as possible. Making use of the above, instead of performing only one cut (parallel to a face) we could perform three cuts (parallel to the three pairs of faces) and then make up a half with the four pieces that satisfy one or all of the three conditions of being 'above', 'in front', or 'on the left' (and the other half with the four pieces satisfying two or none of these conditions). (What is the point of these digressions? They are an attempt to show that phenomena of this kind do not derive from the principles or assumptions of probability theory – in which case one might well have called them 'miraculous'. They may show up, in their own right, in any kind of applications whatsoever. The fact is simply that *the exploitation and the study* of methods based on disorder is more frequent and 'relevant' in probability theory than elsewhere.)

This should give some idea (as well as, in a sense, some of the reasons) of why it is, in complicated situations where some kind of 'disorder' prevails, that something having the appearance of 'order' often emerges.

A further fact, which serves to 'explain' why it is that this 'order generated out of chaos' often has the appearance of a normal distribution, is that out of all distributions having the same variance the normal has maximum entropy (i.e. the minimum amount of information).

Among the discrete distributions with preassigned possible values x_h and prevision $\sum p_h x_h = m$, those which maximize $\sum p_h \,|\log p_h\,|$(the sums $\sum p_h = 1, \sum p_h x_h$ and $\sum p_h x_h^2$ being fixed) are obtained by setting the $\partial/\partial p_h$ of $\sum p_h \{|\log p_h| + Q(x)\}$ equal to 0, where

$Q(x)$ is a second degree polynomial. In other words, we set $-\log p_h + Q(x) = 0$, from which it follows that

$$p_h = \exp(-Q(x)) = K \exp\left\{-\frac{1}{2}(x-m)^2\right\}.$$

If we choose the x_h to be equidistant from each other, and then let this distance tend to zero, we obtain the normal distribution.

In Chapter 3, 3.8.5, we briefly mentioned the idea of information but without going at all deeply into it. In the same way here, without going into the relevant scientific theory, we merely note that, when considering the distribution of velocities in the kinetic theory of gases, 'the same variance' corresponds to 'kinetic energy being constant' (and this may suggest new connections with Maxwell's conclusions; see Section 7.6.7, above).

We could continue in a like manner: there appear to be an endless variety of ways in which the tendency for the normal distribution to emerge occurs.[42]

It is easy to understand the wonder with which its appearance in so many examples of statistical distributions (e.g. in various characteristics of animal species etc.) was regarded by those who first came across the fact, and it is also easy to understand the great, and somewhat exaggerated, confidence in its universal validity that followed.

A typical expression of this mood is found in the following passage of Francis Galton's (it appears in his book *Natural Inheritance*, published in 1889, in the chapter entitled, 'Order in apparent chaos', and the passage is reproduced by E.S. Pearson in one of his 'Studies in the history of probability and statistics' (*Biometrika* (1965), pp. 3–18), which also contains a number of other interesting and stimulating quotations):

> 'I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the "Law of Frequency of Error". The Law would have been personified by the Greeks and deified, if they had known of it. It reigns with serenity and in complete self-effacement amidst the wildest confusion. The huger the mob, and the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of Unreason. Whenever a large sample of chaotic elements are taken in hand and marshalled in the order of their magnitude, an unsuspected and most beautiful form of regularity proves to have been latent all along. The tops of the marshalled row form a flowing curve of invariable proportions; and each element, as it is sorted into place, finds, as it were, a pre-ordained niche, accurately adapted to fit it. If the measurement of any two specified Grades in the row are known, those that will be found at every other Grade, except towards the extreme ends, can be predicted in the way already explained, and with much precision.'

Are statements of this kind acceptable? It seems to me the answer can be both yes and no. It depends more on the nuances of interpretation than on any general principle of whether such statements are correct or not.

[42] A similar 'tendency' for the normal distribution to appear operates, although in a different manner, in problems of statistical inference, as a result of more and more information being acquired. We mention this now merely to make the above survey complete, and in no way to anticipate what will be said later (Chapter 11, 11.4.6–11.4.7, and Chapter 12, 12.6.5).

The idea that all natural characteristics have to be normally distributed is one that can no longer be sustained: it is a question that must be settled empirically.[43] What we are concerned with in the present context, however (and not only in relation to the passage above but also to the numerous other statements, of a more or less similar nature, that one comes across practically everywhere), are the attitudes adopted in response to the 'paradox' of a 'law' governing the 'accidental', which surely obeys no rules.

Perhaps the following couple of sentences will suffice as a summary of the circumstances capable of differentiating and revealing the attitudes which I, personally, would characterize as 'distorted' or 'correct', respectively:

a) there exist chance phenomena which are really under control, in that they follow the 'rules of chance phenomena', and there are others which are even more chancy, accidental in a more extreme sense, irregular and unforeseeable, occurring 'at random', without even obeying the 'laws of chance phenomena';
b) chance phenomena – the completely accidental, those which are to a large extent irregular or unforeseeable, those occurring 'at random' – are those which presumably 'obey the laws of chance phenomena'; these laws express no more and no less than that which can be expected in the absence of any factor which allows one to a large extent to foresee something falling outside the ambit formed by the overwhelming majority of the vast number of possible situations of chaos.

Even when expressed in this way, the two alternatives are very vague (and it would be difficult to avoid this – I certainly did not succeed). They may be sufficient, however, to remove some of the ambiguity from Galton's position, because they show up what the essential ambiguity is that has to be overcome.

Having said this, it remains for me to make clear that I consider (*a*), the *first* interpretation, to be '*distorted*', and (*b*), the *second* interpretation, to be the *correct one*.

The reasons for this are those that have been presented over and over again in the context of concrete problems. There is no need to repeat them here, nor is there any point in adding further general comments or explanations; these, I am afraid, would inevitably remain at a rather vague level.

7.7 Proof of the Central Limit Theorem

7.7.1. We now give the proof of the central limit theorem. This is very short if we make use of the method of *characteristic functions* – although this has the disadvantage of operating with purely analytic entities, having nothing to do with one's intuitive view of the problem. It has the advantage, however, that the very simple proof that can be given for the case of Heads and Tails (confirming something that we have already established in a variety of alternative ways) will turn out to be easily adapted, with very little effort, to provide a proof for very much more general cases.

[43] One must not adopt the exaggerated view that all, or almost all, statistical distributions are normal (a habit which is still widespread, although not so much as it was in the past). Around 1900, Poincaré made the acute observation that 'everyone believes in it: experimentalists believing that it is a mathematical theorem, mathematicians believing that it is an empirical fact'.

For the gain at a single trial of Heads and Tails ($X_i = \pm 1$ with $p_i = \frac{1}{2}$ the characteristic function is given by

$$\frac{1}{2}\left(e^{iu} + e^{-iu}\right) = \cos u.$$

For the sum Y_n of n such trials (stochastically independent summands), we have $(\cos u)^n$. In the standardized form, $Y_n/\surd n$, this becomes $[\cos(u/\surd n)]^n$, with logarithm equal to $n \log \cos(u/\surd n)$.

Since $\log \cos x = -\frac{1}{2}x^2[1+\varepsilon(x)]$ (where $\varepsilon(x) \to 0$ as $x \to 0$), we have $n \log \cos(u/\surd n) = -\frac{1}{2}n(u/\surd n)^2[1+\varepsilon(u/\surd n)] = -\frac{1}{2}u^2[1+\varepsilon(u/\surd n)] \to -\frac{1}{2}n^2$, and, passing from the logarithm back to the characteristic function, we obtain

$$\left[\cos\left(u/\surd n\right)\right]^n \to e^{-\frac{1}{2}u^2} \quad \text{as } n \to \infty. \tag{7.35}$$

This is precisely the characteristic function of the standard normal distribution, and so the theorem is proved.

But this does not merely hold for the case of Heads and Tails. The essential property that has been used in the proof is not the fact that the characteristic function of the individual gain is given by $\varphi(u) = \cos u$, but only that its qualitative behaviour in the neighbourhood of the origin is

$$\log\phi(u) = -\frac{1}{2}u^2\left[1+\varepsilon(u)\right].$$

This requires only that the variance be finite (the value 1 is merely due to the convention adopted previously).

Therefore: *the central limit theorem holds for sums of independent, identically distributed random quantities provided the variance is finite.*[44]

It is clear, however, that the conclusion does not require the distributions to be identical, nor the variances to be equal: given the qualitative nature of the circumstances which ensure the required asymptotic behaviour, purely qualitative conditions should suffice.

It is perhaps best to take one step at a time, in order to concentrate attention on the two different aspects separately. Let us begin with the assumption that the distributions do not vary but that the variances may differ from trial to trial (to be accurate, we should say that the type of distribution does not vary; for the sake of simplicity, we shall continue to assume the prevision to be zero).

[44] If the variance is infinite, the central limit theorem can only hold in what one might call an anomalous sense; that is by not dividing Y_n by $\surd n$, as would be the case for the normal distribution itself, but rather, if at all, through some other kind of standardization procedure, $(Y_n \square A_n)/B_n$, with A and B appropriate functions of n. This holds (see Lévy, *Addition*, p. 113) for those distributions for which the mass outside $\pm x$, if assumed concentrated at these points, has a moment of inertia about the origin which is negligible compared to that of the masses within $\pm x$ (i.e. the ratio tends to zero as $x \to \infty$). These distributions, plus those with finite variances, constitute the 'domain of attraction' of the normal distribution.

There exist other stable distributions (with infinite variances), each having its own domain of attraction (see Chapter 8, Section 8.4).

In other words, we continue to consider the X_i to be independently and identically distributed, standardized random quantities ($\mathbf{P}(X_i) = 0$, $\mathbf{P}(X^2) = 1$), but with the summands X_i replaced by $\sigma_i X_i$ (with $\sigma_i > 0$, and varying with i). Explicitly, we consider the sums

$$Y_n = \sigma_1 X_1 + \sigma_2 X_2 + \ldots + \sigma_n X_n.$$

We again let $\phi(u)$ denote the characteristic function of the X_i and $\varepsilon(u)$ the correction term defined by $\log \phi(u) = -\frac{1}{2}u^2(1+\varepsilon(u))$. The characteristic function of $\sigma_i X_i$ is then given by $\phi(\sigma_i u)$, with

$$\log\phi(\sigma_i u) = -\frac{1}{2}u^2\left[\sigma_i^2 + \sigma_i^2 \varepsilon(\sigma_i u)\right].$$

By taking the product of the ϕ, and the sum of the logarithms, we obtain, for the sum Y_n,

$$\begin{aligned} \log\prod_{i=1}^{n}\phi(\sigma_i u) &= \sum_{i=1}^{n}\log\phi(\sigma_i u) = -\frac{1}{2}u^2\left[\sum_{i=1}^{n}\sigma_i^2 + \sum_{i=1}^{n}\sigma_i^2\varepsilon(\sigma_i u)\right] \\ &= -\frac{1}{2}s_n^2 u^2\left[1 + \sum_{i=1}^{n}\frac{\sigma_i^2\varepsilon(\sigma_i u)}{s_n^2}\right], \end{aligned} \tag{7.36}$$

where s_n^2 denotes the variance of Y_n; that is

$$s_n^2 = \mathbf{P}\left(Y_n^2\right) = \sum_{i=1}^{n}\sigma_i^2.$$

For the standardized Y_n, that is Y_n/s_n, we have (substituting u/s_n for u)

$$-\frac{1}{2}u^2\left[1 + \sum_{i=1}^{n}\frac{\sigma_i^2}{s_n^2}\varepsilon\left(\frac{\sigma_i u}{s_n}\right)\right], \tag{7.37}$$

and, hence, the validity of the central limit theorem depends on the fact that the 'correction term', given by the sum, tends to 0 as $n \to \infty$.

The sum in question is a weighted mean (with weights σ_i^2) of the $\varepsilon(\sigma_i u/s_n)$. Each term tends to 0 as n increases, provided $s_n \to \infty$, because then we will have $(\sigma_i u/s_n) \to 0$. This means that the series formed by summing the variances σ_i^2 must diverge (and this becomes the first condition). This is not sufficient, however. For example, if we took each σ_i very much greater than the previous ones we could make the ratios σ_n/s_n arbitrarily close to 1 and tending to 1, and the correction term would be $\varepsilon(u)$; this would not be improved by dividing u by s_n. The same problem arises if all the ratios, or an infinite number of them, are greater than some given positive number. To ensure that the correction term tends to 0, it is therefore necessary to have $\sigma_n/\mathrm{s}_n \to 0$; this also turns out to be sufficient[45] (and will be the second condition).

[45] This is intuitively obvious but it is perhaps best to give the proof, because it is a little less immediate than it might appear at first sight. Fixing $\varepsilon > 0$, we have $\sigma_n/s_n < \varepsilon$ for all n greater than some given N; each σ_i will therefore satisfy $\sigma_i, < \varepsilon s_i < \varepsilon s_n$ for $n > i > N$, and $\sigma_i < s_i \leqslant s_N$ for $i \leqslant N$. Given that $s_n \to \infty$, for all n greater than some given M we have $s_n > s_N/\varepsilon$, that is $s_N/s_n < \varepsilon$, and hence we have, for $i < N$, also $\sigma_i/s_n < s_i/s_n < s_N/s_n < \varepsilon$.

To summarize: *the central limit theorem holds for sums of independent random quantities whose distributions, apart from the variances,*[46] *are the same, provided the total variance diverges* ($s_n \to \infty$) *and the ratios* $\sigma_n/s_n \to 0$. In other words, the theorem holds if, roughly speaking, the contribution of each term becomes negligible compared with that of the total of the preceding terms.

7.7.2. In particular, this holds for bets on Heads and Tails (or at dice, or other games of chance; trials with $p \neq \frac{1}{2}$) when we allow the stakes, S_i to vary from trial to trial. The individual random gains are $S_i(E_i - p)$, the variance is $\sigma_i = S_i\sqrt{(p\tilde{p})}$, and the standardized random quantity is

$$X_i = (E_i - p)/\sqrt{(p\tilde{p})}$$

(for $p = \frac{1}{2}, \sigma_i = \frac{1}{2}S_i$ and $X_i = 2(E_i - \frac{1}{2}) = 2E_i - 1$, as is always used in the case of Heads and Tails).

In order to fix ideas, we can develop considerations of more general validity in the context of this case; this should clarify the result we have obtained. Recall that the σ_i, are the same as the S_i, apart from a change in the unit of measurement.

If the sum of the σ_i^2 were convergent, it would be like having a sum with a finite number of terms (one could stop when the 'remainder' becomes negligible when modifying the distribution obtained). Not only would the argument used to prove that such a distribution is normal not then be valid any longer, but a different argument would even allow one to exclude it being so (except in the trivial case in which all the summands are normal).[47] The condition $s_n \to \infty$ is therefore necessary.

So far as the condition $\sigma_n/s_n \to 0$ is concerned, notice that it is satisfied, in particular, if the σ_n are bounded above (in the above example, this would be the case if the stakes could not exceed some given value) and that this is the only case in which the conclusion holds independently of the order of summation. Were this not the case, one could, in fact, alter the original order, ($\sigma_1, \sigma_2, \ldots, \sigma_n, \ldots$, in such a way as to every now and again (and hence infinitely often) make the ratio σ_n^2/s_n^2 greater than $\frac{1}{2}$ (say). One possible procedure would be the following: after, say, 100 terms, if the next one (σ_{101}) is too small to give $\sigma_{101}^2/s_{101}^2 > \frac{1}{2}$, insert between the 100th and the 101st the first of the succeeding σ which is $> s_{101} \cdot \sqrt{2}$. Proceed for 100 more terms, and then repeat the operation; and so on.[48]

The conclusion is, therefore, the following: if we have a countable number of summands with no preassigned order, then only the more restrictive condition of *bounded variance* (all the $\sigma_i \leqslant K$) ensures the validity of the central limit theorem (the integers serve as subscripts, but these are merely used by convention to distinguish the summands). On the other hand, if the order has some significance – for example, chronological – then things are different, and the previous conclusion ($s_n \to \infty, \sigma_n/s_n \to 0$) is, in fact, valid and less restrictive.[49]

[46] See the statement of the theorem for the full meaning of this phrase.

[47] By virtue of Cramèr's theorem (Chapter 6, Section 6.12).

[48] There is no magic in the figure 100; it was chosen in this example because it seemed best to have a number neither too large, nor too small. The rule must guarantee that all the terms of the original sequence appear in the rearranged sequence (part of the original sequence might be permanently excluded if at each place one term were chosen on the basis of the exigencies of magnitude, or whatever).

[49] It seems to be important, both from a conceptual and practical point of view, to distinguish the two cases. In general, however (and, indeed, always, so far as I know), it seems that one only thinks in terms of the case of ordered sequences. It is always necessary to ask oneself whether the symbols actually have a genuine meaning.

If, in particular, we wish to consider the case in which all the stakes (and, therefore, the variances) are increasing, the condition means that the σ_n, must increase more slowly than any geometric progression ('eventually'; i.e. at least from some given point on).

The reason for such bounds is also obviously intuitive. In fact, if a very large bet arises, it, by itself, will influence the shape of the distribution in such a way as to destroy the approach to the normal which might have resulted from all the preceding bets.

It remains to consider now what happens if we let not only the σ_i vary, but also the (standardized) distributions of the X_i. All the expressions that we wrote down in the previous case remain unaltered, except that, in place of $\phi(\sigma_i u)$ and $\varepsilon(\sigma_i u)$, we must now write $\phi(\sigma_i u)$ and $\varepsilon(\sigma_i u)$, allowing for the fact that the distribution (and hence the ϕ and ε) may vary with i.

All that we must do, then, is to examine the 'correction term' in the final expression. The single ε is now replaced by the ε_i and, in order to be able to draw the same conclusion, it will be sufficient to require that the $\varepsilon_i(u)$ all tend to zero in the same way as $\mu \to 0$. In other words, it is sufficient that there exists a positive $\varepsilon(u)$ tending to 0 as $\mu \to 0$, which provides an upper bound for the $\varepsilon_i(u)$; $|\varepsilon_i(u)| \leqslant \varepsilon(u)$.

As far as the meaning of this condition is concerned, it requires that (for the standardized summands, X_i) the masses far away from the origin tend to zero in a sufficiently rapid, uniform manner. More precisely, it requires that $\mathbf{P}(|X_i| \geqslant x)$ be less than some $G(x)$, the same for all the X_i which is decreasing and tending to zero rapidly enough for $\int x^2| \mathrm{d}G(x)| < \infty$ (see Lévy, *Addition*, p. 106).

A sufficient condition is that of Liapounov, which is important from a historical point of view in that it provided the basis of the first rigorous proof of the central limit theorem under fairly unrestrictive conditions (1901). The condition requires that, for at least one exponent $2 + \delta > 2$, the moment exists for all the X_i

$$\mathbf{P}\left(|X_i|^{2+\delta}\right) = a_i < \infty \tag{7.38}$$

and that

$$(a_1 + a_2 + \ldots + a_n)/s_n^{2+\delta} \to 0 \quad \text{as } n \to \infty.$$

7.7.3. All that remains is to ask whether the three conditions

$$\left(s_n \to \infty, \sigma_n / s_n \to 0, |\varepsilon_i(u)| < \varepsilon(u)\right)$$

that we know to be sufficient for the validity of the central limit theorem are also necessary. The answer, a somewhat unusual one, perhaps, but one whose sense will become clearer later, is that they are not necessary, but almost necessary.

The necessary and sufficient conditions constitute the so-called Lindeberg–Feller theorem. This improves upon the version which we gave above, and which is known as the Lindeberg–Lévy theorem. The range of questions involved is very extensive and has many aspects, the theory having been developed, more or less independently, and in various ways, by a number of authors, especially in the period 1920–1940. In a certain sense, Lindeberg was the one who began the enterprise, and Lévy and Feller produced the greatest number of contributions (along with Cramèr, Khintchin and many others). Our treatment has looked at just a few of the most important, but straightforward, aspects of the theory. The presentation is original, however, in that we have made an effort to unify everything (arguments, choice of notation and terminology, emphasis on what is fundamental, and what is peripheral), and because of the inclusion of examples and comments, perhaps novel, but, in any case, probably useful, at least for clarification.

This digression, in addition to providing some historical background, serves to give warning of the impossibility of giving a brief, complete clarification of the phrase, 'not necessary, but almost necessary', which constituted our temporary answer. Not only would we need to include even those things which we intended to omit, but we would also need to give the reasons why we intended omitting them. As an alternative, we shall give the gist of the matter, together with a few examples: the gist is that the conditions only need be weakened a little – 'tinkered with' rather than substantially altered. We have already seen the trivial case (summands all normal) which holds without requiring $s_n \to \infty$; the necessary and sufficient conditions are analogous to the necessary ones but refer to the sums, Y_n, rather than to the summands (allowance is therefore made for intuitive cases of compensation among the effects of different summands, or, for an individual summand, of compensation between a large value of σ_i and a very small $\varepsilon_i(u)$, that is an X_i with a distribution which is almost exactly normal).

An extension of a different kind is provided by the following, which seems, for a variety of reasons, worth mentioning: *a sum of X_h with infinite variances can also tend to a normal distribution* (although within pretty narrow confines, and with rather peculiar forms of normalization). The condition (for X_h with the same distribution F and prevision 0) is that $U(a) = \int_{-a}^{a} x^2 \, \mathrm{d}F$ 'varies slowly' as $a \to \infty$ (that is for every $k > 0$ we must have

$$U(ka)/U(a) \to 1, \quad \text{as}\, a \to \infty,$$

although, by hypothesis, $U(a) \to \infty$). This implies, however, that, *for every $\alpha < 2$*, the moments of order α are finite (this involves the same integral as above, but with $|x|^\alpha$ replacing x^2), and that one does not have convergence for the distributions of the $Y_n/\sqrt{n}$ (but instead for some other sequence of constants, to be determined for each case separately). These are the two remarks we made above; note that the second takes up and clarifies the remarks of Chapter 6, 6.7.1, and the first footnote of that section: an example of this is provided by $f(x) = 2|x|^{-3} \log|x|(|x| \geqslant 1)$, where the normalization is given by $Y_n/(\sqrt{n} \log n)$ (see Feller, Vol. II, in several places).

7.7.4. *A complement to the 'law of large numbers'.* This complement (and we present here the important theorem of Khintchin) is included at this point simply for reasons of exposition. In fact, the method of proof is roughly the same as that given above.

We know that for the arithmetic mean, Y_n/n, of the first n random quantities, X_i with $\mathbf{P}(X_i) = 0$, we have $Y_n / n \overset{\cdot}{\to} 0$, and hence $Y_n / n \overset{<}{\to} 0$ (the quadratic and weak laws of large numbers, respectively), provided that the variances σ_i^2 are bounded and have a divergent sum. Khintchin's result states that $Y_n / n \to 0$ also holds if the variances are not finite, provided the X_i all have the same distribution. (Other cases also go through, under appropriate restrictions.)

If $\mathbf{P}(X_i) = 0$, we have $\log \phi(u) = u\varepsilon(u)$, with $\varepsilon(u) \to 0$ as $u \to 0$. For Y_n/n, the logarithm of the characteristic function is therefore

$$n \ . \ u/n \ . \ \varepsilon(u/n) = u\varepsilon(u/n) \to 0 \quad \text{as}\, n \to \infty,$$

and hence the characteristic function tends to $e^0 = 1$, and the distribution to $F(x) = (x > 0)$ (all the mass concentrated at the origin): in other words, the limit of Y_n/n (in the weak sense) is 0; $Y_n/n \overset{<}{\to} 0$.

If the distributions of the X_i are not all equal, the $\phi_i(u)$, and therefore the $\varepsilon_i(u)$, will be different. The logarithm of the characteristic function of Y_n/n will then be equal to $(u/n) \sum\varepsilon_i(u/n) = u \times$ the (simple) arithmetic mean of the $\varepsilon_i(u/n)$. Here, too, it suffices to assume that the $\varepsilon_i(u)$ tend to 0 in the same way; that is for all i we must have $|\varepsilon_i(u)| < \varepsilon(u)$, with $\varepsilon(u)$ positive and tending to zero as $u \to 0$. The condition concerning the distributions of the X_i is similar to the previous one (except that it entails the first moment rather than the second): $\mathbf{P}(|X_i| \geqslant x)$ all bounded by one and the same $G(x)$, decreasing and tending to 0 rapidly enough to ensure that $\int x|\mathrm{d}G(x)| < \infty$. Clearly, this is a much less restrictive condition than the previous one: the present condition concerns the influence of the far away masses on the evaluation of the *prevision* (whereas the variance can be infinite); the previous condition was concerned with the influence on the evaluation of the *variance* (which had to exist, or, perhaps better, to be finite).

In order to understand the Khintchin theorem, it is necessary to recollect that we here assume for $\mathbf{P}(X)$ the value given by $\hat{\mathbf{P}}(X)$ (by virtue of the convention we adopted in Chapter 6, 6.5.6).

The theorem states that the mean Y_n/n of $X_1, X_2, \ldots, X_n, \ldots$ (independent, with the same distribution F), converges in a *strong* way to a constant a if and only if the mean value $F(\square) = \hat{\mathbf{P}}(X)$ of F exists and equals a (weak convergence can hold even without this condition).[50]

This property shows $\hat{\mathbf{P}}$ to have an interesting probabilistic significance and hence to appear as something other than merely a useful convention.

[50] More general results, with simple proofs, are given in Feller, Vol. II (1966), pp. 231–234.