# 4

# Conditional Prevision and Probability

## 4.1 Prevision and the State of Information

We have all at times insisted on making clear the fact that every prevision and, in particular, every evaluation of probability, is conditional; not only on the mentality or psychology of the individual involved, at the time in question, but also, and especially, on the state of information in which he finds himself at that moment.

Those who would like to 'explain' differences in mentality by means of the diversity of previous individual experiences, in other words – broadly speaking – by means of the diversity of 'states of information', might even like to suppress the reference to the first factor and include it in the second. A theory of this kind is such that it cannot be refuted, but it seems (in our opinion) rather meaningless, being untestable, vacuous and metaphysical; in fact, since two different individuals (even if they are identical twins) cannot have had, instant by instant, the same identical sensations, any attempt at verification or refutation assumes an absurd hypothesis. It is like asking whether or not it is true that had I lived in the Napoleonic era and had participated in the Battle of Austerlitz I would have been wounded in the arm.

As long as we are just referring to evaluations relative to the same individual and state of information, there is no need to make any explicit mention of it; for example instead of $\mathbf{P}(E)$, writing something like $\mathbf{P}(E|H_0)$, where $H_0$ stands for 'everything that is part of that individual's knowledge at that instant'. Indeed, something which in itself is so obvious, and yet so complicated and vague to put into words, is clearer if left to be understood implicitly rather than if one thinks of it condensed into a symbol, like $H_0$.

Naturally, things change if we want to combine previsions that are relative to different states of information, and we shall see later that one cannot do without this. In precise terms, we shall write $\mathbf{P}(E|H)$ for the *probability 'of the event E conditional on the event H'* (or even the *probability 'of the conditional event $E|H$'*), which is the probability that You attribute to $E$ if You think that in addition to your present information, that is the $H_0$ which we understand implicitly, *it will become known to You that H is true (and nothing else)*. This $H$, on the other hand, may be a combination of 'simpler' events (this is obvious, but it is better to point it out explicitly); in other words, it can denote, in a condensed manner, a whole complex of new information, no matter how extensive (so long as it is well delimited).

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

The above explanations may be useful as a preliminary guide to the meaning of the concept of *conditional probability*, **P**($E|H$) – and, more generally, of *conditional prevision*, **P**($X|H$) – which we are about to introduce. We ought to warn the reader, however, against an overhasty acceptance of these initial explanations, which, of necessity, skipped over certain important details, a discussion of which would have been premature (see the *Remarks* given in Chapter 11, 11.2.2). Think, instead, in terms of the definition that we are now going to give.

The definition is based on the same concepts and criteria that we met previously (see Chapter 3), except for the additional assumption that *any agreement made* – that is any *bet* or *penalty clause* – will remain *without effect if H does not turn out to be true*: in other words, everything is *conditional on the 'hypothesis' H*. (Concerning the terminology 'hypothesis', see Section 4.4.2.)

The 'first criterion' provides an intuitive explanation, which we exploit only to anticipate the meaning of the 'theorem of compound probabilities'. By paying the price **P**($HE$), I can be sure of receiving one lira if $HE$ occurs; but I can obtain the same result by paying **P**($E|H$) only if I know $H$ is true, and I can arrange for this amount, $S$ = **P**($E|H$), in the case of the occurrence of $H$ by paying $S$. **P**($H$) now; hence

$$\mathbf{P}(HE)=\mathbf{P}(H)\cdot\mathbf{P}(E\,|\,H). \tag{4.1}$$

The same is true if, instead of an event $E$, I consider an arbitrary random quantity $X$; it is sufficient to observe that $HX$ coincides with $X$, or is zero, depending on whether $H$ is true or false, and the extension of the preceding argument to this case becomes obvious.

## 4.2 Definition of Conditional Prevision (and Probability)

In order to give definitions of *conditional probability* and *conditional prevision*, and as a foundation for rigorous proofs, we choose to base ourselves on the 'second criterion'.

***Definition.*** Given a random quantity $X$ and a *possible* event $H$, suppose it has been decided that You are subject to a penalty

$$L=H\left(\frac{X-\bar{x}}{k}\right)^2$$

($k$ fixed arbitrarily in advance), where $\bar{x}$ is the value which You are at liberty to choose as You like. (Note: we have $L$ = 0 if $H$ = 0 = *false*; $L=\left[(X-\bar{x})/k\right]^2$ if $H$ = 1 = *true*.)

**P**($X|H$), *the prevision of X conditional on H* (in your opinion), is the value $\bar{x}$ that You choose for this purpose.

In particular, if $X$ is an event, $E$, then **P**($E|H$), so defined, is called *the probability of E conditional on H* (in your opinion).

***Coherence.*** It is assumed that (in normal circumstances) You do not prefer a given penalty if You can choose a different one which is *certainly* smaller.

*A necessary and sufficient condition* for coherence in the evaluation of **P**($X|H$), **P**($H$) and **P**($HX$), is compliance with the relation

$$\mathbf{P}(HX)=\mathbf{P}(H).\mathbf{P}(X\,/\,H), \tag{4.2}$$

in addition to the inequalities $\inf(X|H) \leq \mathbf{P}(X|H) \leq \sup(X|H)$, and $0 \leq \mathbf{P}(H) \leq 1$; in the case of an event, $X = E$, relation (4.1),

$$\mathbf{P}(HX) = \mathbf{P}(H).\mathbf{P}(E \mid H),$$

is called the *theorem of compound probabilities*, and the inequality for $\mathbf{P}(X|H)$ reduces to $0 \leq \mathbf{P}(E|H) \leq 1$ (being = 0, or = 1, in the case where $EH$, or $\tilde{E}H$, respectively, is impossible).

By $\inf(X|H)$ and $\sup(X|H)$, we denote the lower and upper bounds of the possible values for $X$ which are *consistent* with $H$; such values are simply the possible values of $HX$, with the proviso that the value 0 is to be included only if $X = 0$ is compatible with $H$ (i.e. if $HX$ can come from $H = 1$, $X = 0$, and not only, as is necessarily the case, from $H = 0$, with $X$ arbitrary).

## 4.3 Proof of the Theorem of Compound Probabilities

Let us consider first the case of events, and denote by $x, y, z$ the values we suppose to be chosen, according to the given criterion, as evaluations of $\mathbf{P}(E|H)$, $\mathbf{P}(H)$, $\mathbf{P}(HE)$. In this case, the theorem is expressed by (4.1), and, with the above notation, it states that $z = xy$.

The penalty (taking the coefficient $k = 1$) turns out to be

$$L = H.(E - x)^2 + (H - y) + (HE - z)^2,$$

that is, in the three cases to be distinguished,

$$HE(H = E = HE = 1), \quad H\tilde{E}(H = 1, E = HE = 0)$$
$$\text{and } \tilde{H}(H = HE = 0),$$

we have

$$\begin{aligned} HE:&\quad L = u = (1 - x)^2 + (1 - y)^2 + (1 - z)^2 \\ H\tilde{E}:&\quad L = v = x^2 + (1 - y)^2 + z^2 \\ \tilde{H}:&\quad L = w = y^2 + z^2 \end{aligned}$$

Geometrically (interpreting $x, y, z$ as Cartesian coordinates) (Figure 4.1), the penalties $u$, $v$, $w$, in the three cases, are the squares of the distances of the point $(x, y, z)$ from, respectively, the point $(1, 1, 1)$, the point $(0, 1, 0)$, and the $x$-axis (that is from the point $(x, 0, 0)$, the projection of $(x, y, z)$ onto the axis). The four points lie in the same plane if a fifth one, $(x, 1, z/y)$, does also (this is the intersection of the line joining the last two with the plane $y = 1$), and this therefore must coincide with $(x, 1, x)$ – which is on the line joining the first two points. In order for this to happen, we must have $z = xy$, that is the point $(x, y, z)$ must lie on this paraboloid (and, of course, inside the unit

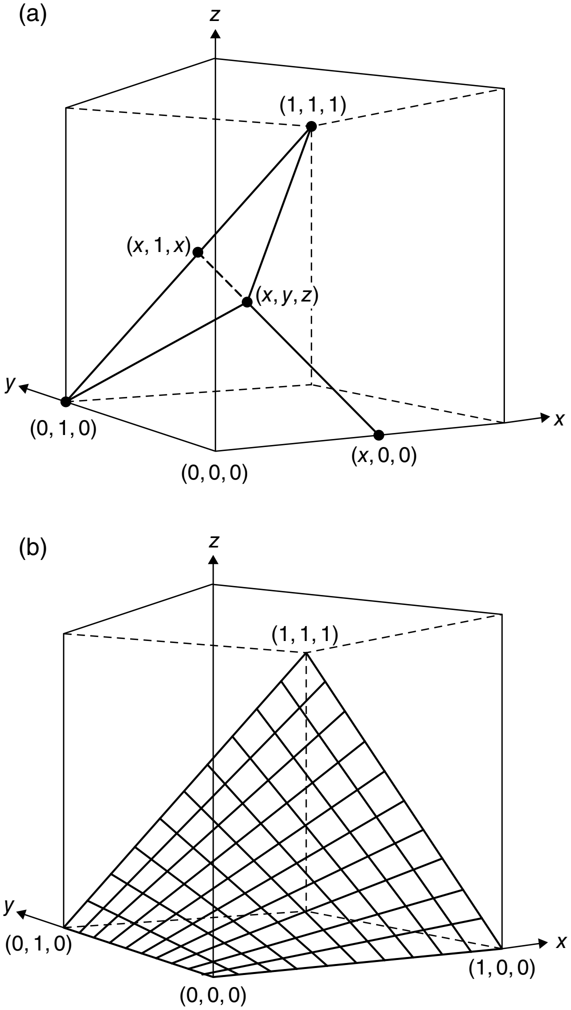

**Figure 4.1** The two diagrams illustrate, in two stages, the argument given in Section 4.3: (a) shows why the prevision-point $(x, y, z)$ must lie on a generator of the paraboloid $z = xy$ (presenting visually the argument of the text); (b) shows the set of all possible prevision-points (the part of the paraboloid inside the unit cube).

cube): in this case, it is not possible to simultaneously shorten the three distances; in other cases this is possible.[1]

Turning to the general case of an arbitrary random quantity $X$, let us again use the notation $x = \mathbf{P}(X|H)$, $y = \mathbf{P}(H)$ and $z = \mathbf{P}(HX)$, and observe that the previous representation is still valid, except that, instead of the two points (1, 1, 1) and (0, 1, 0) on the line

[1] A more detailed discussion can be found in B. de Finetti, 'Probabilità composte e teoria delle decisioni', *Rendic. di Matematica* (1964), 128–134. An English translation of this appears in B. de Finetti, *Probability. Induction and Statistics*, John Wiley & Sons (1972).

$y = 1$, $z = x$, we must consider all the points whose abscissae $x$ are possible for $X$ and compatible with $H$. In fact, expanding (in canonical form) we have,

$$L = H.(X-x)^2 + (H-y)^2 + (HX-z)^2$$
$$= H\left[(X-x)^2 + (1-y)^2 + (X-z)^2\right] + (1-H)\left(y^2+z^2\right).$$

If $(x, y, z)$ were not on the paraboloid $z = xy$ (i.e. not in the plane through the line $y = l$, $z = x$ and the point $(x, 0, 0)$), one could, as before, make it approach, simultaneously, both the $x$-axis and each point of the given line. In order that this should not be possible, it is necessary, in addition, to restrict oneself to the area (a quadrilateral bounded by the straight lines generating the paraboloid) given by

$$0 \le y \le 1 \text{ and } \inf(X \mid H) \le x \le \text{sub}(X \mid H).$$

The convenience of substituting $y = 0$ and $y = 1$ for any $y < 0$, or $y > 1$, respectively, is obvious; that $x$ must not be outside the bounds for $(X|H)$ becomes clear (without spending time on the calculations) if one observes, in mechanical terms, that in order to cancel out a force acting at the point $(x, y, xy)$ directed towards $(x, 0, 0)$ – that is tending to make it approach the $x$-axis – it is necessary to have a force directed towards $(x, l, x)$, which is opposite (or, alternatively, more than one, directed towards points which are on both sides of this point on the line $y = 1$, $z = x$). If the possible points were all on one side (and only in this case) all distances could be shortened by moving towards the nearest bound.[2]

## 4.4 Remarks

4.4.1. Let us note first of all that, as we have already seen in passing, in questions concerning the conditioned event, $E|H$, the event $E$ itself does not actually enter the picture: the cases to be distinguished are, in fact, $HE$, $H\tilde{E}$, $\tilde{H}$. Since $H$ is called the 'hypothesis' of the conditioned event, $HE$ could be called the 'thesis', $H\tilde{E}$ the 'antithesis', and $\tilde{H}$ the 'antihypothesis'. Every conditioned event $E|H$ could then be written in the *reduced* form 'thesis'| 'hypothesis', $HE|H$ (in fact, it does not matter whether one bets that if $H$ occurs $E$ does, or that if $H$ occurs both $H$ and $E$ do). One might consider $E|H$ as a tri-event with values $1|1 = 1$, $0|1 = 0$, $0|0 = 1|0 = \emptyset$, where $1 =$ *true*, $0 =$ *false*, $\emptyset =$ *void*, depending on whether it leads to a *win* or a *loss* or a *calling off* of a possible conditional bet. More generally, for a conditioned (random) quantity, $X|H$, one could put $X|1 = X$, $X|0 = \emptyset$ (if $\emptyset$ is thought of as outside the real field, $\inf(X|H)$ and $\sup(X|H)$ automatically acquire the desired meaning, introduced previously as a convention). The systematic use of algorithms based on this set of ideas does not seem sufficiently worthwhile to compensate for the bother of introducing them; however, this brief mention may suggest a few arguments for which it might turn out to be suitable.

[2] This conclusion might fail to hold if the possible points were all on the same side of $(x, 1, x)$, but having this point as a bound (lower or upper). We will dwell upon detailed considerations of this kind in the sequel.

4.4.2. As far as the use of the term 'hypothesis' for $H$ is concerned, it should be unnecessary to point out that it refers only to the position of $H$ in $E|H$ (or in $X|H$), and that, apart from this, $H$ is any event whatsoever. We say this merely to avoid any possible doubts deriving from memories of obsolete terminologies (like 'probability of the hypotheses' or, even worse, 'of the causes', a notion charged with metaphysical undertones.

4.4.3. This being so, together with $E|H$ one can always consider $H|E$ as well (where $E$ becomes the 'hypothesis'); indeed, since $EH = HE$, we obtain immediately the relationship between the probabilities of these two conditional events:

$$\mathbf{P}(EH) = \mathbf{P}(E)\mathbf{P}(H|E) = \mathbf{P}(H)\mathbf{P}(E|H),$$

which implies that

$$\mathbf{P}(E|H) = \mathbf{P}(E)\frac{\mathbf{P}(H|E)}{\mathbf{P}(H)}\left(\text{provided}\,\mathbf{P}(H) \neq \mathbf{0}\right); \tag{4.3}$$

this last formula is *Bayes's theorem*, whose fundamental rôle will be seen over and over again. Observe, however, that it is merely a different version, or corollary, of the theorem of compound probabilities.

The fact that relationships of this kind are of interest, also shows why it is not convenient (contrary to appearances) to consider systematically the reduced form, $HE|H$ (i.e. $E|H$ with $E\tilde{H} = 0$), which would simply give

$$\mathbf{P}(E|H) = \mathbf{P}(E)/\mathbf{P}(H).$$

4.4.4. Anyway, on the basis of the theorem of compound probabilities, one can deduce (provided $\mathbf{P}(H) \neq 0$) that

$$\mathbf{P}(E|H) = \mathbf{P}(HE)/\mathbf{P}(H); \tag{4.4}$$

this shows that, from a formal standpoint, and assuming coherence, conditional probability is not a new concept, since it can be expressed by means of the concept of probability that we already possess. This observation is, in fact, made use of in the axiomatic treatments; however, using this approach, one obtains the formula, not the meaning. For this reason (and also so as not to leave out the case, albeit a limit-case, where $\mathbf{P}(H) = 0$) we have considered it necessary to start from the *essential* definitions and *prove* the *theorem* of compound probabilities (instead of reducing it to a definition, which could appear arbitrary).

## 4.5 Probability and Prevision Conditional on a Given Event *H*

4.5.1. Let us examine how, for all the events $E$, and random quantities $X$, of interest, one passes from probabilities $\mathbf{P}(E)$, and previsions $\mathbf{P}(X)$ (we will call them *actual*, in order to distinguish them), to those *conditional* on a given event $H$. We already know that $\mathbf{P}(E|H) = \mathbf{P}(HE)/\mathbf{P}(H)$ – let us suppose that $\mathbf{P}(H) \neq 0$ – and, in general, that $\mathbf{P}(X|H) = \mathbf{P}(H|X)/\mathbf{P}(H)$, but it is useful to think about this and give some illustrations, and in the meantime to observe also that $\mathbf{P}(\cdot|-H)$ is additive, etc.; that is it is an admissible **P** (an element of P in the linear ambit we started with). In fact,

$$\mathbf{P}(X+Y\,|\,H)=\mathbf{P}(HX+HY)/\mathbf{P}(H)=\mathbf{P}(HX)/\mathbf{P}(H)+\mathbf{P}(HY)/\mathbf{P}(H);$$

in particular, for events $A$ and $B$, $\mathbf{P}(A + B|H) = \mathbf{P}(A|H) + \mathbf{P}(B|H)$ and in the case of incompatibility the same holds for $\mathbf{P}(A \vee B|H)$; we therefore have

$$\mathbf{P}(\tilde{E}\,|\,H)=\mathbf{1}-\mathbf{P}(E\,|\,H),\ \text{and so on.}$$

4.5.2. Decomposing $E$ into $EH + E\tilde{H}$ (incompatible parts, constrained to be in $H$ and in $\tilde{H}$, respectively) one sees immediately that it is the first part which gives rise to the value $\mathbf{P}(E|H) = \mathbf{P}(EH)/\mathbf{P}(H)$ (i.e. it increases in the same ratio as $\mathbf{P}(H)$ to 1, and the same is true for $H$, which goes from $\mathbf{P}(H)$ to $\mathbf{P}(H|H) = 1$), whereas the contribution of the second part is zero

$$\mathbf{P}(E\tilde{H}\,|\,H)=\mathbf{P}(E\tilde{H}H\,|\,H)=\mathbf{P}(0\,|\,H)=0).$$

Interpreting the events as sets, and the probability as mass, one obtains for this case a more effective and instructive image; considering the probability conditional on $H$ implies:

- making all masses outside the set $H$ (‘hypothesis’) vanish,
- normalizing the remaining masses (i.e. altering them, proportionately, so that the total mass is again ‘one’).

The same rule holds for $\mathbf{P}(X|H)$, and could also be interpreted within this same framework (but in a less obvious form and, for the time being anyway, unintuitively.

4.5.3. Mentioning this is not only convenient from the point of view of having the rule of calculation easily at hand but, as we have said, it is conceptually instructive. If these obvious considerations are well understood, confusions that are often irremediable will be avoided. The acquisition of a further piece of information, $H$ – in other words, *experience*, since experience is nothing more than the acquisition of further information – acts always and only in the way we have just described: *suppressing the alternatives that turn out to be no longer possible* (i.e. leading to a more strict limitation of expectations). As a result of this, the probabilities are the $\mathbf{P}(E|H)$ instead of the $\mathbf{P}(E)$, but *not because experience has forced us to modify or correct them, or has taught us to evaluate them in a better way* (even if statements of this kind might perhaps appear tolerable at the level of a crude popularization): the probabilities are the same as before – even if in complicated cases this is less evident and perhaps, at first sight, not even believable – *except for the disappearance of those which dropped out and the consequent normalization of those which remained.*

## 4.6 Likelihood

4.6.1. *Bayes's theorem* – in the case of events $E$, but not random quantities $X$ – permits us to write $\mathbf{P}(\cdot|H)$ in the form we met above, a form which is often more expressive and practical:

$$\mathbf{P}(E\,|\,H)=\mathbf{P}(E)\mathbf{P}(H\,|\,E)/\mathbf{P}(H)=K.\mathbf{P}(E)\mathbf{P}(H\,|\,E), \tag{4.5}$$

where the normalizing factor, $1/\mathbf{P}(H)$, can be simply denoted by $K$, and, more often than not, can be obtained more or less automatically without calculating $\mathbf{P}(H)$. For this reason, it is often convenient to talk simply in terms of *proportionality* (i.e. by considering $\mathbf{P}(\cdot|H)$ only up to an arbitrary, nonzero, multiplicative constant, which can be determined, if necessary, by normalizing).

One could say that $\mathbf{P}(\cdot|H)$ is proportional to $\mathbf{P}(\cdot)$ and to $\mathbf{P}(H|\cdot)$, where the dot stands for $E$, thought of as varying over the set of all the events of interest. More concisely, this is usually expressed by saying that

$$\text{'final probability'} = K \ \text{'initial probability'} \times \text{'likelihood'}$$

where $= K$ denotes proportionality, and we agree to call: the *initial* and *final* probabilities those not conditional or conditional on $H$, respectively (i.e. evaluated before and after having acquired the additional knowledge in question, $H$), and the *likelihood* of $H$ given $E$, the $\mathbf{P}(H|E)$ thought of as a function of $E$ (and possibly multiplied by any factor independent of $E$, e.g. $1/\mathbf{P}(H)$, the use of which would allow the substitution of ' = ' for '= $K$', or anything resulting from the omission of common factors, more or less cumbersome, or constant, or dependent on $H$). The term 'likelihood' is to be understood in the sense that a larger or smaller value of $\mathbf{P}(H|E)$ corresponds to the fact that the knowledge of the occurrence of $E$ would make $H$ either more or less probable (our meaning would be better conveyed if we spoke of the 'likelihoodization' of $H$ by $E$).

4.6.2. This discussion leads to an understanding of how it should be possible to pass from the initial probabilities to the final ones through intermediate stages, under the assumption that we obtain, successively, additional pieces of information $H_1, H_2, \ldots, H_n$ (giving, altogether, $H = H_1H_2 \ldots H_n$). In fact, one can also verify analytically that

$$\begin{aligned}\mathbf{P}(E|H_1H_2) &= \mathbf{P}(EH_1H_2)/\mathbf{P}(H_1H_2)\\ &= \left[\mathbf{P}(E)\mathbf{P}(H_1|E)\mathbf{P}(H_2\,|\,EH_1)\right]/\left[\mathbf{P}(H_1)\mathbf{P}(H_2|H_1)\right]\\ &= K.\mathbf{P}(E).\mathbf{P}(H_1|E).\mathbf{P}(H_2\,|\,EH_1)\end{aligned}$$

= (the probability of E) × (the likelihood of $H_1$ given $E$
× (the likelihood of $H_2$ given $EH_1$).

In general,

$$\begin{aligned}\mathbf{P}(E\,|\,H) &= \mathbf{P}(E\,|\,H_1H_2\ldots H_n)\\ &= K\,.\mathbf{P}(E).\mathbf{P}(H_1\,|\,E).\mathbf{P}(H_2\,|\,EH_1).\mathbf{P}(H_3\,|\,EH_1H_2)\\ &\quad \ldots\mathbf{P}(H_n\,|\,EH_1H_2\ldots H_{n-1}).\end{aligned}$$

Although the introduction of the term 'likelihood' merely gives a name to a factor in Bayes's formula, which refers to its rôle in the formula (in addition to the existing term, conditional probability, and apart from the indeterminacy we agreed to by defining it up to multiplicative factors), it has the advantage of emphasizing this factor, which will be present in various forms in more and more complicated problems.

## 4.7 Probability Conditional on a Partition $\mathscr{H}$

Let us consider a (finite[3]) partition $\mathscr{H} = (H_1, H_2,\dots, H_S)$, and the probabilities, $\mathbf{P}(E|H_j)$, of an arbitrary event $E$ conditional on each of the $H_j$. Since $EH_1 + EH_2 + \dots + EH_S = E(H_1 + H_2 + \dots + H_s) = E.1 = E$, and $\mathbf{P}(EH_j) = \mathbf{P}(H_j)\mathbf{P}(E|H_j)$, one has

$$\mathbf{P}(E) = \sum_j \mathbf{P}(H_j)\mathbf{P}(E|H_j): \tag{4.6}$$

in words, it is the weighted average, with weights $\mathbf{P}(H_j)$, of the probabilities of $E$ conditional on the different $H_j$. In particular, it lies between them:

$$\min\mathbf{P}(E|H_j) \le \mathbf{P}(E) \le \max\mathbf{P}(E|H_j) \tag{4.7}$$

(and it coincides with them if they are all equal). We shall call this property (which is not always valid for infinite partitions) the *conglomerative property* of conditional probability (and prevision).

If we consider as a random quantity, and denote by $\mathbf{P}(E|\mathscr{H})$, the quantity whose value is $\mathbf{P}(E|H_1)$ if $H_1$ occurs, and so on, in other words, in formulae,

$$\begin{aligned}\mathbf{P}(E|\mathscr{H}) &= H_1\mathbf{P}(E|H_1) + H_2\mathbf{P}(E|H_2) + \dots + H_S\mathbf{P}(E|H_S)\\ &= \sum_{H\in\mathscr{H}} H.\mathbf{P}(E|H),\end{aligned} \tag{4.8}$$

we can write the expression above as

$$\mathbf{P}(E) = \mathbf{P}\big[\mathbf{P}(E|\mathrm{H})\big]. \tag{4.9}$$

More generally, we have, of course,

$$\mathbf{P}(X) = \mathbf{P}\big[\mathbf{P}(X|\mathrm{H})\big]. \tag{4.10}$$

The procedure displayed above, obtaining a prevision by decomposing it into previsions conditional on the alternatives in a partition (which may often be chosen in such a way as to make the task easier, either through mathematical convenience, or through psychological judgement), is very helpful in many cases. We shall see this in ad hoc examples, and even more so in the frequent references we make to it in what follows.

## 4.8 Comments

The idea of considering $\mathbf{P}(E|\mathscr{H})$ as a random quantity requires some further comment.

4.8.1 As we have said, a random quantity $X$ is a quantity that is well defined, in an objective sense, although unknown. Does this mean then that, taking $X = \mathrm{P}(E|\mathscr{H})$ with the meaning that $X = \mathrm{P}(E|H_j) = x_j$ if $H_j$ occurs, under such a hypothesis it is objectively true that the value of the above-mentioned probability is $x_j$? Certainly not; but the

[3] This restriction cannot be removed without further conditions (see later: Section 4.19).

possibility of this doubt must be removed. The problem is meaningful only after a particular evaluation of the probabilities $\mathbf{P}(E|H_j)$ has been taken into consideration; whether this is a subjective evaluation of a given individual, or a hypothetical evaluation. Given this, independently of the fact that the $x_j$ have been determined as a result of these actual or hypothetical evaluations, instead of by measuring magnitudes or by choosing them at random, they are objectively determined numbers. That the value of $X$ turns out to be $x_j$ when $H_j$ occurs is true in the sense that $x_j$ is the value that by definition has been associated with $H_j$ The fact that the association is as an evaluation of $\mathbf{P}(E|H_j)$, made at a certain moment, by a certain individual, may or may not be of interest, but is irrelevant to the definition.

4.8.2. For equation 4.9 (or 4.10) to be true, it is of course necessary that $\mathbf{P}$ always refers to the same individual: the average of the $\mathbf{P}_1(E|H_j)$ of one individual weighted by the $\mathbf{P}_2(H_j)$ of another does not give the $\mathbf{P}(E)$ of either of them; neither $\mathbf{P}_1(E)$ nor $\mathbf{P}_2(E)$.

4.8.3. The idea of considering $\mathbf{P}(E|\mathscr{H})$ as a random quantity often leads to a temptation that one should be warned against: this is the temptation of saying that we are faced with an '*unknown* probability', which is either $x_1$ or $x_2$ ... or $x_s$ but we do not know which is the *true* value, $x_j$, until we know which of the hypotheses $H_j$ is the *true* one. At any moment, the probability is that relative to the information one has; it can refer, for convenience, to different hypothetical pieces of information that can be arbitrarily chosen in an infinite number of ways, thus obtaining an infinite number of different conditional probabilities. None of them, and likewise none of the possible hypotheses, has any special status entitling them to be regarded as more or less 'true'. Any one of them could be 'true' if one had the information corresponding to it; in the same way as the one corresponding to one's present information is true at the moment.

4.8.4. In those cases in which it turns out to be *convenient* to refer to a partition – and these are the only cases in which the temptation meets needs which are essentially meaningful – it is a question, as we have just made clear above, of 'probabilities conditional on unknown objective hypotheses'. As usual, by 'convenient' we are referring to making an evaluation easier by taking one step at a time, and by choosing the easiest steps.

Probability is the result of an evaluation; it has no meaning until the evaluation has been made and, from then on, it is known to the one who has made it.[4] For this obvious reason alone, the phrase 'unknown probabilities' is already intrinsically improper, but what is worse is that the improper terminology leads to a basic confusion of the issues involved (or reveals it as already existing). This is the confusion that consists in thinking that the evaluation of a probability can only take place in a certain 'ideal state' of information, in some *privileged* state; in thinking that, when our information is different (as it will be, in general), more or less complete, in part more so, in part less so, or different in kind, we should abandon any probabilistic argument (and, perhaps, rely on adhockeries).

4.8.5. On the contrary, there are innumerable possible partitions, which might appear more or less special in character. In order to restrict ourselves to a single example, let us

[4] *For me, someone else's* evaluation may be unknown, etc.; however, it is for me an objective fact (an evaluation), independently of the subjective reasons which, within him, have led to its determination.

assume that we have to make a drawing from an urn containing 100 balls. We do not know the respective numbers of white and black balls but, for the sake of simplicity, let us suppose that we attribute equal probabilities to symmetric compositions, and equal probability to each of the 100 balls: the probability of drawing a white ball is therefore $= \frac{1}{2}$. Someone might say, however, that the *true* probability is not $\frac{1}{2}$ but $b/100$, where $b$ denotes the (unknown) number of white balls: the true probability is thus unknown, unless one knows how many white balls there are. Another person might observe, on the other hand, that 1000 drawings have been made from that urn and, happening to know that a white ball has been drawn $B$ times, one could say that the *true* probability is $B/1000$. A third party might add that both pieces of information are necessary, as the second one could lead him to deviate slightly from attributing equal probabilities to all the balls (accepting it, in the absence of any facts, as a frequency, somewhat divergent from the actual composition). A fourth person might say that he would consider the knowledge of the position of each ball in the urn at the time of the drawing as constituting complete information (in order to take into account the habits of the individual doing the drawing; his preference for picking high or low in the urn): alternatively, if there is an automatic device for mixing them up and extracting one, the knowledge of the exact initial positions which would allow him to obtain the result by calculation (emulating Laplace's demon).[5]

Only in this case (given the ability) would one arrive, at last, at the true, special partition, which is the one in which the theory of probability is no longer of any use because we have reached a state of certainty. The probability, 'true but unknown', of drawing a white ball is 100% under the hypothesis that the ball to be drawn is white, and 0% under the hypothesis that it is black.

But uncertainty is what it is; information is the information that one actually has (until we can obtain more, and so reduce uncertainty). If one wants to make use of the theory of probability one can only apply it to the actual situation; if one wants to make a plaything of it, little problems can be invented on which it is imagined that one can pin the label 'objective' in a facile fashion; one must not mix up the two things, however: even Don Quixote did not consider venturing forth upon the world astride a rocking-horse.

## 4.9 Stochastic Dependence and Independence; Correlation

4.9.1. The probability of $E$ conditional on $H$, $\mathbf{P}(E|H)$, can be either equal to $\mathbf{P}(E)$, or greater, or less. This means that the knowledge (or the assumption) that $H$ is true either does not change our evaluation of probability for $E$, or leads us to increase it, or to diminish it, respectively. In the first case, one says that $E$ is *stochastically independent* of $H$ (or *uncorrelated* with $H$); in the other cases, $E$ is said to be *stochastically dependent* on $H$; more precisely, either *positively* or *negatively correlated* with $H$.

We observe straightaway that the property is symmetrical: the theorem of compound probabilities enables us to write down immediately (for $\mathbf{P}(E)$ and $\mathbf{P}(H)$ nonzero)

$$\frac{\mathbf{P}(E|H)}{\mathbf{P}(E)} = \frac{\mathbf{P}(H|E)}{\mathbf{P}(H)} = \frac{\mathbf{P}(EH)}{\mathbf{P}(E).\mathbf{P}(H)'} \tag{4.11}$$

[5] In practice, the various partitions which may present themselves as 'reasonable' are, in fact, much more numerous than in this example, which is already quite 'traditional' in itself.

and hence it turns out that *the ratio by which the probability of E increases or decreases when conditioned on H is the same as that for H conditioned on E, and it is also equal to the ratio between the probability of EH and the product of the probabilities of E and H.* Obviously, in the case of stochastic independence, this product is **P**(*EH*); in fact,

$$\mathbf{P}(EH)=\mathbf{P}(H).\mathbf{P}(E\,|\,H)=\mathbf{P}(H)\mathbf{P}(E) \quad \text{assuming} \quad \mathbf{P}(E\,|\,H)=\mathbf{P}(E). \tag{4.12}$$

Therefore, we may also say, in a symmetric form, that two events *are* stochastically independent (uncorrelated) or are negatively or positively correlated (with each other). It is clear that if $E$ and $H$ are positively correlated the same is true for $\tilde{E}$ and $\tilde{H}$, whereas the reverse is true for $E$ and $\tilde{H}$, and for $\tilde{E}$ and $H$: if one of the pairs is stochastically independent (uncorrelated) the same is true in all four cases. (Verify this as an exercise.)

*Remarks.* This symmetry in behaviour between *positive correlation* and *negative correlation* no longer holds, however, when more than two events are considered. Although positive correlations, however strong, are always possible, negative correlations are not possible unless they are very weak (at least on average), the more so the greater the number of events.

The proof will be given (for the general case of random quantities) in Section 4.17.5: at the present time we do not even have the concepts required to express the statement, except in the informal way given above. At this juncture, it is necessary to point out the conceptually significant aspects of the matter rather than leaving it until the technical exposition to which we referred. In that exposition, Figures 4.3a and 4.3b reveal the reason, in an intuitive fashion, by means of the following analogy: *it is possible to imagine as many vectors as we wish forming arbitrarily small angles, but not forming angles which are all 'rather' obtuse*:[6]

4.9.2. For more than two events, $E_1, E_2, \ldots, E_n$, say, we could, of course, consider pairwise stochastic independence, $\mathbf{P}(E_iE_j) = \mathbf{P}(E_i)\mathbf{P}(E_j)$, $i \neq j$, but, in fact, they are termed *stochastically independent* only if

$$\mathbf{P}\left(E_{i_1}E_{i_2}\ldots E_{i_k}\right)=\mathbf{P}\left(E_{i_1}\right)\mathbf{P}\left(E_{i_2}\right)\ldots\mathbf{P}\left(E_{i_k}\right) \tag{4.13}$$

holds for any arbitrary product of the events E$_i$: this condition is, as we shall see later, more restrictive. This property, if it holds for the $E_i$ also holds if some of them are replaced by their negations $\tilde{E}_i$ as we have already observed in the case of two events. We therefore have, *for stochastically independent events $E_i$*, whose probabilities are denoted by $p_i$, that the probability of a product, such as $\tilde{E}_1\tilde{E}_2E_3E_4\tilde{E}_5$, is obtained by simply writing $p$ in place of $E$; thus $\tilde{p}_1\tilde{p}_2p_3p_4\tilde{p}_5$, that is $(1 - p_1)(1 - p_2)p_3p_4(1 - p_5)$. More generally, *for any event E, which is logically dependent on the $E_i$, and expressed arithmetically in*

[6] This sentence is rather vague, but rather than make it complicated it is preferable to ask the reader to accept it for now, simply as a reference to what we shall see in more detail shortly.

*terms of them in canonical form* (*with* $+, .$ and $\sim$), *the probability is expressible in terms of the* $p_i$ *by the same formula.*[7] For example, if

$$E = (E_1 \vee E_2E_3)(\tilde{E}_4 \vee E_5\tilde{E}_6),$$

expanding, we obtain

$$E = (E_1 + E_2E_3 - E_1E_2E_3)(\tilde{E}_4 + E_5\tilde{E}_6 - \tilde{E}_4E_5\tilde{E}_6),$$

and so on, and finally one could substitute $p$ for $E$. In fact, since no $E$ appears repeated in both parentheses, we can substitute straightaway (without arriving at a single sum of products) and write

$$\mathbf{P}(E) = (p_1 + p_2p_3 - p_1p_2p_3)(\tilde{p}_4 + p_5\tilde{p}_6 - \tilde{p}_4p_5\tilde{p}_6).$$

4.9.3. A particular, celebrated case, and one which has been extensively studied, is that of *stochastically independent and equally probable events*, $p_i = p$; this is the *Bernoulli scheme*, also referred to as that of 'repeated trials'. For every $E$, logically dependent on $n$ such events, the probability $\mathbf{P}(E)$ turns out to be expressed by a polynomial in $p$ (of degree at most $n$); for example, the $E$ considered above (depending on the six events $E_1 \ldots E_6$) would have the probability

$$\begin{aligned}\mathbf{P}(E) &= (p + p^2 - p^3)(\tilde{p} + p\tilde{p} - p\tilde{p}^2) = p(1 + p - p^2)(1 - p + p^2 - p^3)\\ &= p - p^3 + p^4 - 2p^5 + p^6.\end{aligned}$$

Less obvious algebraically, but more meaningful, would be the analogous expression as a homogeneous polynomial of degree $n$ in the two variables $p$ and $\tilde{p} = (1 - p)$; it is obtained, in an obvious fashion, by multiplying each term by a suitable power of $(p + \tilde{p}) = 1$. In the previous example, operating in the two factors right from the beginning, one has, for example,[8]

$$\begin{aligned}\mathbf{P}(E) &= p\left[(p + \tilde{p})^2 + p(p + \tilde{p}) - p^2\right]\tilde{p}\left[(p + \tilde{p})^2 + p(p + \tilde{p}) - p\tilde{p}\right]\\ &= p\tilde{p}^5 + 5p^2\tilde{p}^4 + 9p^3\tilde{p}^3 + 8p^4\tilde{p}^2 + 2p^5\tilde{p}.\end{aligned}$$

[7] The reduction to canonical form is not necessary: it is only required to draw attention to the fact that, when we expand, powers $E_i^k$, with $k > 1$, do not appear formally; to these would correspond probabilities $p_i^k$ instead of $p_{\mathrm{i}}$ as must be the case by virtue of the idempotence of the $E_{\mathrm{i}}$, $E_i^k$. For example, if $E = (E_1 \vee E_2)(E_1 \vee E_3) = (E_1 + E_2 - E_1E_2)(E_1 + E_3 - E_1E_3)$ and we substituted straightaway, we would wrongly obtain $\mathbf{P}(E) = (p_1 + p_2 - p_1p_2)(p_1 + p_3 - p_1p_3) = p_1^2\tilde{p}_2\tilde{p}_3 + p_i(\tilde{p}_2p_3 + p_2p_3) + p_2p_{3,}$ whereas, in place of the first factor, $p_1^2$, we should have $p_1$. As a general rule, one might consider substituting the $p_i$ for the $E_i$, suppressing the exponents at the end: this procedure could be dangerous, however, since if the $p_i$ were equal, for example, and were replaced straightaway by $p$, one would make a mistake in the opposite direction.

[8] By introducing the *ratio*, $r = p/\tilde{p}$ (see Chapter 5), we have $p^h\tilde{p}^{n-h} = \tilde{p}^n r^h$, and therefore the polynomial in $p$ and $\tilde{p}$ can be written as $\tilde{p}^n \times$ a polynomial in $r$; in the example given, we would have $\mathrm{P}(E) = \tilde{p}^6(r + 5r^2 + 9r^3 + 8r^4 + 2r^5)$.

The significance of this lies in the following: the coefficients denote the number of constituents of $E$ corresponding to the different frequencies of the $E_i$. In precise terms, the coefficient of $p^h \tilde{p}^{n-h}$ is the number of constituents in which $h$ of the $E_i$ occur, and $n - h$ do not: in other words, with $h$ factors of the form $E_i$ and $n - h$ of the form $\tilde{E}_i$. In the example given, one sees that there is one constituent with a single occurrence (i.e. $(1 \vee 0.\ 0)\ (\tilde{0} \vee 0.\tilde{0})$), five with two, nine with three, eight with four and two with five (this is easily verified because the two factors each have five favourable constituents, of which those containing 0, 1, 2, 3 occurrences number, respectively, 0, 1, 3, 1 and 1, 2, 2, 0).

4.9.4. An even more special case is that in which $p = \frac{1}{2}$ This is usually referred to as the case of *Heads and Tails* (although we could also think in terms of any other interpretation and application, and although the case of Heads and Tails is an exceptional one, where some 'objective circumstance' forces us to adopt this evaluation of probability). In this case, each constituent has probability $p^h \tilde{p}^{n-h} = \left(\frac{1}{2}\right)^n$ and $\mathbf{P}(E) = \left(\frac{1}{2}\right)^n$ × the sum of the coefficients of the polynomial in $p$ and $\tilde{p}$ (or in $r$), which is, in other words, the ratio between the number of constituents (or cases) which are favourable to $E$, and the total number ($2^n$) of constituents.

## 4.10 Stochastic Independence Among (Finite) Partitions

4.10.1 There is an obvious and immediate extension of the notion of stochastic independence from the case of events to that of (finite) partitions; in other words, if one wants to use such terminology, to multi-events, like $E' = \left(E_1', E_2', \ldots, E_{m'}'\right)$ and $E'' = \left(E_1'', E_2'', \ldots, E_{m''}''\right)$, and, in particular, to random quantities with a finite number of possible values. It will simply imply that every event of a partition is stochastically independent of every event of the other one: $\mathbf{P}\left(E_h' E_k''\right) = \mathbf{P}\left(E_h'\right)\mathbf{P}\left(E_h''\right)$ ($h$ = 1, 2,..., $m'$; $k$ = 1, 2,..., $m''$), and, in particular, for random quantities $X$ and $Y$ it will mean that

$$\mathbf{P}\left[(X,Y) = (x_h, y_k)\right] = \mathbf{P}\left[(X = x_h).(Y = y_k)\right] = \mathbf{P}(X = x_h).(Y = y_k). \tag{4.14}$$

And so on for three or more partitions or random quantities (referring always to the finite case).

4.10.2. Let us now prove that pairwise stochastic independence is, as we said, a necessary but not sufficient condition for the stochastic independence of $n$ events (and, *a fortiori*, of $n$ partitions): two examples will suffice.

Let $A$, $B$, $C$, $D$ be the events of a partition, to each of which we attribute probability $\frac{1}{4}$ The events $E_1 = D + A$, $E_2 = D + B$, $E_3 = D + C$ are pairwise independent ($E_iE_j = D$, $\mathbf{P}(E_iE_j) = \frac{1}{4}$ are $\mathbf{P}(E_i)\mathbf{P}(E_j) = \frac{1}{2}\cdot\frac{1}{2}$), but are not so when taken three at a time, since $E_1E_2E_3 = D$, and the probability of the product of all three of them is still $\frac{1}{4}$ instead of $\frac{1}{8}$.

Similarly, considering $A + B$, $B + C$, $C + A$, the products two at a time would have probability $\frac{1}{4}$, but the product of all three is impossible and therefore has probability zero and not $\frac{1}{8}$.

More generally, one can have stochastic independence up to a given order, '$m$ by $m$' say, but riot beyond this, as the following example (a generalization of the previous ones) shows. Let $E_1$, $E_2$,..., $E_m$ be stochastically independent events each of probability $\frac{1}{2}$ (i.e. every 'constituent' has probability $\left(\frac{1}{2}\right)^m$), and let $E$ be the event which consists of the fact that among the $E_i$ there are an odd number of false ones: $E = \left(\tilde{E}_1 + \tilde{E}_2 + \ldots + \tilde{E}_m = \text{odd}\right)$. It is clear that $E$ is logically dependent on the $E_i$ (by definition, and, on the other hand, $EE_1 \ldots E_m = 0$ with certainty, since either some of the $E_i$ are 0, or all of the $\tilde{E}_i$ and their sum are 0, hence not odd, so that $E = 0$), but is stochastically independent of $m - 1$ of them (conditionally on any results of these, $E$ coincides either with the omitted event or with its negation).

4.10.3. Suppose we have two partitions, into $m'$ events $E_1^{'} \ldots E_{m'}^{'}$ and into $m''$ events $E_1^{''} \ldots E_{m''}^{''}$, respectively. To say that in each of them the probabilities of the different events are equal (to $p' = 1/m'$ and $p'' = 1/m''$, respectively) and that they are stochastically independent, implies that the $m = m'm''$ events $E_h^{'} E_k^{''}$ of the product-partition all have the same probability, $p = p'p'' = 1/(m'm'') = 1/m$; conversely, this property implies the two previous ones. The same obviously holds for three or more partitions. We shall come back to this fact, which is the basis for many applications of the combinatorial type.

4.10.4. If we have different partitions, or multi-events, which are stochastically independent and have equally distributed probability (e.g. successive drawings with replacement from an urn, with fixed probabilities of drawings for balls of $m$ different colours, $p_1 + p_2 + \ldots + p_m = 1$), we have an extension of the Bernoulli scheme given above; 'repeated trials' for multi-events. It is clear how the considerations made in the previous case could be generalized: for every event $E$ which is logically dependent on $n$ $m$-events, the probability $\mathbf{P}(E)$ can be expressed as a polynomial $\sum c_{h_1 h_2 \ldots h_m} p_1^{h_1} p_2^{h_2} \ldots p_m^{h_m}$ (the sum being over all $m$-tuples of non-negative integers with sum = $n$). The coefficients give the number of favourable constituents containing the $i$th result $h_i$ times ($i = 1, 2, \ldots, m$). In the case of equal probabilities ($p_1 = p_2 = \ldots p_m = 1/m$), a generalization of Heads and Tails ($m = 2$), the probabilities are

$$\begin{aligned}\mathbf{P}(E) &= (1/m^n) \times \text{the sum of the coefficients of the polynomial}\\ &= \text{the ratio of the number of constituents (or cases)}\\ &\quad\ \text{favourable to } E \text{ and the total number} (m^n) \text{of}\\ &\quad\ \text{all constituents (possible cases).}\end{aligned} \tag{4.15}$$

## 4.11 On the Meaning of Stochastic Independence

4.11.1. It is absolutely essential to continue to underline the fact that the notion of stochastic independence does not belong to the domain of the logic of certainty, but to that of prevision, and that therefore – like probability and prevision – it has a *subjective* meaning. After presenting the necessary details in an abstract setting, we shall need to dwell upon the various considerations required to illustrate them in practice. This is of paramount importance if one takes into account that people usually seem to think – or, at least, allow it to be thought, since objections are rarely put forward – that the meaning of stochastic independence is self-evident and objective, and that this property always holds, except for special cases of interdependence. So much so that in applications to many practical problems[9] one often comes across notions and formulae that are valid if the hypothesis of stochastic independence is adopted, but where this hypothesis does not turn out to be justified and is not, in fact, introduced explicitly, but only tacitly, and perhaps inadvertently. The habit of simply saying 'independence', as if it were a unique notion, plays a part in obscuring the special nature of the notion of stochastic independence. For the sake of brevity, we shall also adopt this habit when there is no ambiguity, or when it is not required to underline the sense: we shall only do it, however, after having given warning of this, and of the existence of other notions which are, in a certain sense, similar. We have already met those of *linear* and *logical* independence (whose meaning resides within the logic of certainty), and the notion of things being *uncorrelated* (which, in the case of events, is synonymous with pairwise stochastic independence, but which, in the case of random quantities, will turn out to be different, as we shall shortly see).

4.11.2. The definition of stochastic independence depends on the evaluation of probability; that is on the choice of a particular **P**. If $A$ and $B$ are two *logically independent* events, an individual can evaluate $\mathbf{P}(A)$, $\mathbf{P}(B)$ and $\mathbf{P}(AB)$ in any way whatsoever, provided that (see Chapter 3, 3.9.4) $\mathbf{P}(AB)$ turns out to be not less than $\mathbf{P}(A) + \mathbf{P}(B) - 1$, and not greater than either of $\mathbf{P}(A)$ and $\mathbf{P}(B)$ (which, in any case, are all numbers between 0 and 1). The ratio $\mathbf{P}(AB)/\mathbf{P}(A)\mathbf{P}(B)$ can, therefore, assume all non-negative values, depending on the appraisal of the person making the evaluation.[10]

Even if, for the sake of brevity, we shall occasionally say that two events (or partitions, etc.) *are* stochastically independent, it must be remembered that this is 'with respect to a given **P**'; in other words, 'according to the opinion of the person who has chosen the evaluation **P**' is to be understood. In particular, in the case of *logically independent* events or partitions, however the probabilities are evaluated, the evaluation extended on the basis of the hypothesis of independence is coherent. If, on the other hand, *we do not have logical independence*, that is some product is impossible, for example $E = E_i' E_j'' E_h'''$ (three elements of three partitions), we necessarily have $\mathbf{P}(E) = 0$: we can have the relation $\mathbf{P}(E) = \mathbf{P}(E_i')\mathbf{P}(E_i'')\mathbf{P}(E_h''')$ if at least one of the

[9] As H. Bühlmann observes (in a report at the ASTIN Congress in Trieste, 1963), the condition of independence is often understood and assumed to be valid when it is not valid at all. He refers to the field of insurance and actuarial mathematics (but what he says is unfortunately true in many other fields). Sometimes, rather than tacitly stating, or considering as obvious, the condition of independence, one considers that 'not knowing much about the interdependence' provides a justification for it. This is tantamount to saying that if we do not know much about the behaviour of a function we can argue as if we knew that it were a constant.

[10] After having evaluated $\mathbf{P}(A) = a$ and $\mathbf{P}(B) = b$, the ratio $\mathbf{P}(AB)/\mathbf{P}(A)\mathbf{P}(B)$ can still assume all non-negative values if $a + b \leq 1$, and all values not less than $1 - (\tilde{a}\tilde{b}/ab)$ otherwise. In any case, the three cases of positive, zero and even negative correlation (since this minimum is always less than 1) remain possible.

factors is zero, the relations $\mathbf{P}\left(E|E_i'E_j''\right)=\mathbf{P}\left(E_h''' \mid E_i'E_j''\right)=\mathbf{P}\left(E_h'''\right)$ (and similar ones) only if all the factors are zero. In other words, the given arithmetic conditions of stochastic independence *cannot hold*, except in the limit cases mentioned above, which do not fall within the definition given in the form of a product, and the more extreme cases, which do not even fall within the definition given in terms of conditional probability. Rather than accept this anomaly, it is preferable to eliminate it by including logical independence as a prerequisite for the definition of stochastic independence. The justification of this is that it is equivalent to taking into account the difference between possible events to which zero probability is attributed and impossible events. This is the same distinction as that between empty sets and nonempty sets of measure zero; a much more fundamental distinction than that between nonempty sets with zero or nonzero measure.

Given these considerations about limit-cases, we can now say (in the case of finite partitions) that *stochastic independence* presupposes *logical independence* (but certainly not vice versa). As far as *linear* dependence is concerned, we recall that it is a particular form of logical dependence and, therefore, it excludes stochastic independence.

In order to complete this hierarchy of notions, let us say at this point that absence of correlation will be a subjective notion weaker than stochastic independence (but when applied under more and more restrictive conditions it may lead to it).

## 4.12 Stochastic Dependence in the Direct Sense

Let us now illustrate some of the kinds of factors that may often influence our judgments of whether events are stochastically independent or dependent. It is necessary to learn how to think carefully about the presence of these factors in order to avoid assuming too readily the hypothesis of stochastic independence, a practice we have already criticized. In putting forward these few cases, we are not attempting an exhaustive treatment, and the mention of these cases is not meant to correspond to a classification having any theoretical value (indeed, the distinctions which we shall make, with the sole aim of drawing together a few examples, might become empty, nebulous abstractions if taken too seriously).

Anyway, without any intention of becoming theoretical, let us call, informally, stochastic dependence *in the direct sense*, the case that arises in the most evident form, and in the most obvious and common examples in treatments from all conceptual viewpoints. This is the case in which the occurrence of an event changes the circumstances surrounding the occurrence of another one (in a way considered relevant to the evaluation of the probability). Standard examples are: drawings from an urn without replacement (where the drawing of a white ball decreases the percentage of white balls for the next drawing); contagious diseases (where a diseased individual increases the probability that people close to him catch the illness); the breakdown of machines and so on (where the difficulties caused by a breakdown of one of them precipitates the breakdown of others); the outcomes of successive trials in a competition (where, due to the initial results, the objective conditions for the succeeding trials change; for example the height of the bar in a high jump competition), and so on.

Examples of this kind draw attention to dependence ‘in one direction’ – chronologically (dependence of what happens afterwards on what has happened before). This corresponds to the interpretation – often, in fact, referred to when considering cases of this kind – based on the idea of ‘*cause*’. That this is irrelevant is seen by observing that the relationship of dependence or independence is symmetric. Anyway, we take this opportunity of remarking that, for ‘conditional’ bets too, it is of no importance whether the ‘fact’ refers to the future or the past and, in particular, whether, chronologically, it follows or precedes the other ‘fact’ assumed as the hypothesis for the validity of the bet. One could very well bet on the occurrence of a certain event today, stipulating that the bet will be effective only if some other event takes place in a month’s time.

Our desire to discuss this case of ‘direct’ dependence was not so much because it needed attention drawing to it, but, on the contrary, to make the reader subsequently aware of the incompleteness of discussions which mention only this form of dependence, and lead one to believe that, apart from such cases, there is no reason to depart from the formulation in terms of stochastic independence. We therefore proceed now to consider certain other examples.

## 4.13 Stochastic Dependence in the Indirect Sense

By this we mean, in an informal way, as above, those cases in which the occurrence of an event has no influence on the occurrence of another one, but in which there are some circumstances that can influence both events. In other words – if one wishes to speak in terms of ‘causes’ – there is a ‘cause’ common to these events, but there is no direct ‘causal’ relationship between them. For example, in considering (the possibility of) two ships both being wrecked in the same area, on the same day (even without assuming collisions or any direct interference of this kind), one might rightly imagine a positive correlation, since both probabilities are influenced in the same way by common circumstances (like the state of the sea; calm or stormy). The same holds true for the deaths of two individuals during next winter, since, if it is very cold, the probability of death will increase for both of them. In the same way, if we ask whether two participants in a competition will achieve better results than some other participant, the result obtained by the latter will influence the two events in the same way, even if one judges the three results to be stochastically independent. This latter example can also be given an interpretation in terms of a game of chance in which $A$ and $B$ ‘win’ if they obtain a greater score than the ‘bank’ does. Interpreting the score as that obtained by throwing a die, then, in terms of the ‘score’ obtained by the ‘bank’, the probabilities of wins for $A$ or $B$, or both, are given by

| the ‘bank’s’ score ($H$): | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $\mathbf{P}(A\|H) = \mathbf{P}(B\|H) =$ | 5/6 | 4/6 | 3/6 | 2/6 | 1/6 | 0 |
| $\mathbf{P}(AB\|H) =$ | 25/36 | 16/36 | 9/36 | 4/36 | 1/36 | 0 |

and averaging (assuming that each of the six cases has probability = 1/6)

$$\mathbf{P}(A) = \mathbf{P}(B) = 15/36 = 5/12 = 41\cdot 67\%$$
$$\mathbf{P}(A)\mathbf{P}(B) = 25/144 = 75/432 = 17\cdot 36\%$$
$$\mathbf{P}(AB) = 55/216 = 110/432 = 25\cdot 45\% > \mathbf{P}(A)\mathbf{P}(B).$$

This example shows that conditional on each of the possible hypotheses for the 'bank's' score, $H$ = ('points' = $h$) with $h$ = 1, 2,..., 6, the two events are stochastically independent, but that this independence conditional on each event of a partition *does not imply stochastic independence*. We will return shortly to an explicit consideration of this notion and this result, to which the case of indirect dependence essentially reduces.

There is one case, however, which derives even less from 'objective' circumstances.

## 4.14 Stochastic Dependence through an Increase in Information

If it is true (as it is, in fact) and if one can justify (as we have, for the moment, simply assumed) that the probability of an event is often evaluated on the basis of observed frequencies of more or less similar events, then this fact implies a stochastic dependence. In fact, observed events provide a certain amount of experience capable of modifying, as time goes on, the evaluations of probabilities based on frequencies. Indeed, it is precisely the analysis based on these present considerations that will lead later (Chapter 11) to an explanation of why and under what conditions such a criterion of evaluation turns out to be justified.

The situation to which we refer is obviously relevant in the case of 'new' phenomena; that is those about which there is little past experience: think, for instance, of the success or failure of the first space launches; of the first trials employing a new drug, or something of that kind; of the probability of death in a species of animal never before observed; of the risks attached to nuclear experimentation, and so on. Putting on one side the hypothesis of 'new', the situation does not change in essence but does change quantitatively, as a few, or even many, trials cannot produce any substantial alteration of a frequency arrived at after a great many previous trials. This is so unless one is led to behave as if faced with a 'new' phenomenon: thinking, for instance, that because of a change in circumstances (or for whatever other reason) the future frequency of an 'old' phenomenon (like mortality, fire, hail, or anything else) will closely resemble the frequency suggested by a small number of recent experiences, rather than the frequency observed in a large number of less recent experiences.[11]

In a certain sense, the situation is the same as that of drawings with replacement from an urn of unknown composition: the probabilities of white balls at successive drawings turn out to be interdependent because the results, as they are obtained, make one's ideas about the composition of the urn more precise (and the smaller the past experience, the greater the influence it has on our ideas). This case could really have been included among the previous examples of indirect dependence (dependence on the

[11] This is the problem studied by American actuaries under the heading of 'Credibility Theory'; see the two lectures by A.L. Mayerson and B. de Finetti containing information and discussion about this topic: *Giorn. Ist. Ital. Attuari* (1964).

unknown composition of the urn); the only difference – an irrelevant one – is the fact that here the composition is an unknown but pre-existent datum, whereas in the other examples we were dealing with the influence of future events, uncertain at the moment when the question was posed. Instead, in the given examples of 'new phenomena' our disposition to review the evaluation was not attributed to ignorance of circumstances, or of specific, objectively determined magnitudes, but, in a general way, to a lack of familiarity with the phenomenon. There may be those who would like to say that such an 'objective magnitude' is the 'constant, but unknown, probability'. We have explained many times, however, that it is not admissible to speak in this way, and we shall also see that it is unnecessary, because, by arguing in a sensible way about meaningful notions, one comes to the same conclusions as would be obtained by meaningless arguments, introducing meaningless notions. Anyway, this means that none of the cases present any essential differences, neither conceptually nor mathematically, notwithstanding the external differences which required us to look at them separately in order to avoid an over-restricted view.

The temptation to proceed further with these considerations, which could not be completed here, is best resisted: we recall that their purpose was simply to persuade the reader that, *in a certain sense, it is stochastic independence which constitutes a rather idealized limit-case*, and that dependence is the norm, rather than the contrary (whose acceptance is the bad habit referred to by Bühlmann; see Section 4.11.1, footnote).

## 4.15 Conditional Stochastic Independence

4.15.1. In the previous examples, we have encountered the notion of conditional stochastic independence (conditional on an event, on a partition); it is necessary to add something more systematic in this connection.

We shall say that $E_1 \dots \mathrm{E}_n$ are stochastically independent with respect to $H$ (or with respect to each $H = H_j$ of a partition) if they are such with respect to the function (or in general the functions) $\mathbf{P}$ of the type $\mathbf{P}(\cdot) = \mathbf{P}(\cdot|H)$ (i.e. $\mathbf{P}(E_1E_2|H) = \mathbf{P}(E_1|H) \cdot \mathbf{P}(E_2|H)$, etc.).

In the example (of beating the 'bank' when throwing dice), we found that $A$ and $B$, stochastically independent with respect to a partition, turned out to be positively correlated; $\mathbf{P}(AB) > \mathbf{P}(A)\mathbf{P}(B)$. We now want to examine the question in general, beginning with a very simple example (less restrictive than the previous one, in the sense that the probabilities of the two events are not assumed to be equal). Let us consider just two hypotheses, $H$ and $\tilde{H}$, with probabilities $c$ and $\tilde{c}$; let the events $A$ and $B$ have probabilities $a'$ and $b'$ conditional on $H$, and $a''$ and $b''$ conditional on $\tilde{H}$. The probability of $AB$ will be

$$\mathbf{P}(AB) = c \cdot \mathbf{P}(AB\,|\,H) + \tilde{c} \cdot \mathbf{P}(AB\,|\,\tilde{H}) = ca'b' + \tilde{c}a''b'', \tag{4.16}$$

whereas, in order that $A$ and $B$ be independent, it should have been

$$\begin{aligned}\mathbf{P}(AB) &= \mathbf{P}(A)\cdot\mathbf{P}(B) = (ca' + \tilde{c}a'')(cb' + \tilde{c}b'')\\ &= c^2a'b' + c\tilde{c}(a'b'' + a''b') + \tilde{c}^2a''b'';\end{aligned}$$

the difference is

$$\begin{aligned}\mathbf{P}(AB)-\mathbf{P}(A)\mathbf{P}(B)&=\left(c-c^2\right)a'b'-c\tilde{c}\left(a'b''+a''b'\right)+\left(\tilde{c}-\tilde{c}^2\right)a''b''\\&=c\tilde{c}\left(a'b'+a''b''-a'b''-a''b'\right)=c\tilde{c}\left(a'-a''\right)\left(b'-b''\right).\end{aligned}\tag{4.17}$$

One therefore has stochastic independence only in the trivial cases: $c = 0$ or 1, or $a' = a''$, or $b' = b''$; in other words, if the two hypotheses do not have zero probability, only if $A$ (or $B$) is stochastically independent of them:

$$\mathbf{P}(A)=\mathbf{P}(A\,|\,H)=\mathbf{P}(A\,|\,\tilde{H}).$$

If this does not happen, one has positive or negative correlation according to whether the probabilities of $A$ and $B$ vary in the same or the opposite sense when conditional on $H$ rather than $\tilde{H}$. This is what we would have expected.

4.15.2. The same problem, with a partition into $s$ hypotheses $H_1 \dots H_s$ instead of two, with probabilities $c_1 \dots c_s$, and with

$$\mathbf{P}(A\,|\,H_j)=a_j,\quad \mathbf{P}(B\,|\,H_j)=b_j,$$

gives:

$$\begin{aligned}\mathbf{P}(A)=a=\sum c_ja_j,\ \ \mathbf{P}(B)=b=\sum c_jb_j,\ \ \sum c_j=1,\\ \mathbf{P}(AB)=\sum c_ja_jb_j=\sum c_j\left[a+\left(a_j-a\right)\right]\left[b+\left(b_j-b\right)\right]\\ =ab+\sum c_j\left(a_j-a\right)\left(b_j-b\right),\\ \mathbf{P}(AB)-\mathbf{P}(A)\mathbf{P}(B)=\sum c_j\left(a_j-a\right)\left(b_j-b\right).\end{aligned}\tag{4.18}$$

One can easily see directly from this expression that if when the $a_j$ increase the $b_j$ increase as well, the difference is positive; that is $A$ and $B$ turn out to be positively correlated (negatively if the change is in the opposite direction): this generalizes the previous conclusion. In particular, if (conditional on each $H_j$) $A$ and $B$ have equal probabilities, $a_j = b_j$, they are positively correlated (so that the conclusion of the example concerning the die and the bank was necessary, not just incidental). More generally, once we have defined correlation between random quantities, we shall see that the expression obtained above will correspond to the following statement: $A$ and $B$ are positively or negatively correlated, or uncorrelated, according to the sense in which the random quantities $X = \mathbf{P}(A|\mathscr{H})$ and $Y = (B|\mathscr{H})$ are correlated; in other words, according to whether $\mathbf{P}(XY) \gtreqless \mathbf{P}(X)\mathbf{P}(Y)$.

4.15.3. The case of conditional stochastic independence gives rise to a particularly interesting case of inductive argument; that is of determining the probabilities of the different possible hypotheses conditional on the information regarding the outcomes of any events which are judged to be *stochastically independent of each other, conditionally on each of the above mentioned 'hypotheses'.*

This is – to refer to the standard example of the classical variety – the case of drawings with replacement from an urn of unknown composition: the hypotheses are the different compositions of the urn (e.g. percentages of white and black balls), the events are the drawing of a white ball on given trials. On the other hand, in order to demonstrate the importance of this in less academic examples, this is often the form of argument used to evaluate the probability of the two hypotheses of the guilt or innocence of an accused man on the basis of the ascertainment of a certain number of facts having the status of 'circumstantial evidence', or 'proof'. If the latter facts differ as much as possible they can, therefore, be taken as stochastically independent of each other, conditional on both hypotheses, and with different probabilities conditionally on the two hypotheses.

It goes without saying that jurors and magistrates would reject with horror the idea of a verdict as an evaluation of probability: in order to have their feet on solid ground, they feel obliged to present as the 'truth', or as a 'certainty', some version which, through the procedures provided, has qualified as the official and compulsory version (and which, therefore, cannot be open to correction, even if an individual who was officially murdered many years ago shows up looking very much alive[12]). It is sad, to say the least, to see such an unconscientious preference for a 'certainty', which is almost always fictitious, rather than a responsible and accurate evaluation of probability. Perhaps the saddest thing, however, is the thought that the world will probably remain for quite some time at the mercy of a mentality so distorted and arrogant that it neither retracts nor hesitates even when faced with the most grotesque absurdities.[13]

One more example: Heads and Tails using a coin that we think may be 'imperfect' (i.e. it may 'favour' one side more than the other). As different 'hypotheses' in this case, one often considers the 'hypothesis of an imperfection giving rise to a probability $p$ of heads', a different 'hypothesis' for each value of $p$, or for a certain number of values $p_h$; for example, in order to simplify matters, increments of 1%. This formulation is not very satisfactory because the definition of a hypothesis on the basis of an evaluation of probability is a nonsense; however, before seeing (in Chapter 11) the way in which an equivalent, and correct, formulation can be given, based on the notion of 'exchangeable events', without speaking of such 'hypotheses', one can accept this image, for the time being, as a 'temporary formulation'. This is acceptable on account of the above observation that it is equivalent in its actual conclusions to the correct formulation, even if it is, strictly speaking, meaningless.

4.15.4. Formally, the particular case we are referring to reduces to the obvious simplification introduced in the expression for $\mathbf{P}(E|H)$ (given in Section 4.6.2), if the items of information $H_i$, which make up $H$, are stochastically independent of each other conditional on the events $E$. Then, in fact, $\mathbf{P}(H_2|EH_1)$ reduces to $\mathbf{P}(H_2|E)$, $\mathbf{P}(H_3|EH_1H_2)$ reduces to $\mathbf{P}(H_3|E)$, and so on, and, finally, the likelihood for the information $H_1\,H_2\,\ldots\,H_n$

[12] As happened recently in Sicily.

[13] Some even assert that in the absence of proofs sufficient for conviction the accused should always be discharged 'for not having committed the crime'. On the other hand, it can well happen that it is certain that one of two suspects is guilty, e.g. one or other, or both, of a married couple (like in the 'Bebawi case', Rome 1966). Judicial wisdom, which ignores common sense, and, therefore, probability, would then have to assert, in effect, that all the inhabitants of the world are under suspicion apart from two people, one of whom is the murderer, who are officially free and protected from any possibility of suspicion.

*Translators' note.* The Bebawis were a married couple appearing in a murder trial, who were each accusing the other of the murder. They were both acquitted on the grounds that the cases against them were insufficiently proved.

(the product of the $H_i$) is nothing other than the product of the likelihoods for the single $H_i$, so that:

$$\mathbf{P}(E\,|\,H)=\mathbf{P}(E\,|\,H_1H_2\ldots H_n)=K\mathbf{P}(E)\mathbf{P}(H_1\,|\,E)\mathbf{P}(H_2\,|\,E)\ldots\mathbf{P}(H_n\,|\,E). \tag{4.19}$$

In a form which is sometimes more expressive, given two events $E$ ($E_h$ and $E_k$, say) we can write

$$\frac{\mathbf{P}(E_h\,|\,H)}{\mathbf{P}(E_k\,|\,H)}=\frac{\mathbf{P}(E_h)}{\mathbf{P}(E_h)}\cdot\frac{\mathbf{P}(H_1\,|\,E_h)}{\mathbf{P}(H_1\,|\,E_k)}\cdot\frac{\mathbf{P}(H_2\,|\,E_h)}{\mathbf{P}(H_2\,|\,E_k)}\cdots\frac{\mathbf{P}(H_n\,|\,E_h)}{\mathbf{P}(H_n\,|\,E_k)}. \tag{4.19′}$$

In other words: the ratio of the final probabilities (of any two events $E$) is given by the ratio between their initial probabilities times the ratios of the likelihoods for each item of information $H_j$. One should note the particular case in which, in place of $E_k$, we substitute the negation $\tilde{E}_h$ of $E_h$: put more succinctly, $E_h = E$ and $E_k = \tilde{E} = 1 - E$, and then one obtains a relationship between the initial and final ratios $\mathbf{P}(E)/\mathbf{P}(\tilde{E})$, and the ratios $\mathbf{P}(H_j|E)/\mathbf{P}(H_j|\tilde{E})$, which we might call *ratios of probability* and *ratios of likelihood*, respectively: we shall talk about this explicitly in Chapter 5, 5.2.4–5.2.5.

This result expresses – at least in the Bayesian version[14] – the 'Likelihood Principle':

'*For the purpose of inferences concerning the events E, the information obtained from the occurrence of the $H_j$ can be arrived at from the knowledge of the likelihoods* $\mathbf{P}(E_h|Hj)$ (*or of their ratios*).'

It is, however, necessary (in order to avoid possible misunderstandings) to underline that this is true *only if* the conditions specified above hold; we will discuss this in greater detail in Chapter 11.

In the meantime, let us point out a qualitative and expressive formulation of one particular conclusion that corresponds to many practical situations:

'*Suppose a thesis* (*e.g. the guilt of an accused man*) *is supported by a great deal of circumstantial evidence of different forms, but in agreement with each other; then even if each piece of evidence is in itself insufficient to produce any strong belief the thesis is decisively strengthened by their joint effect*.'

This statement is known as 'Cardinal Newman's principle', since it was he (taking it over from previous authors) who made it famous as the basis of his mode of argument in his work the 'Grammar of Assent'.

4.15.5. *Remarks*. In the case of *independence* also we find ambiguity, as already illustrated in Section 4.8. There, it was a question of considering as the 'true' probability not that relative to the actual state of information, but a different one, unknown, conditional on some idealized form of unacquired information. Here, it is a question of calling 'independent' those events that are such conditional on a certain 'ideal' partition. Again, a typical example is that of drawings from an urn of unknown composition, which are independent conditional on the knowledge of the composition (or on any assumption

[14] The reservation expressed by this parenthetical clause is due to the fact that some people believe that the sense in which this 'principle' is understood by non-Bayesian authors, and in particular by Allan Birnbaum who has written about it and supported it, is different. Thus far, I have been unable to discover what these supposed essential differences are (apart from the interpretation; subjectivistic or nonsubjectivistic).

about it), but are not independent for someone ignorant of the composition.[15] Precisely because of the interdependence induced by this ignorance, the successive information about the outcomes of the drawings serves to modify the evaluations of probability (in the sense of Section 4.14). In the case of independence, all such information would, by definition, have no effect.[16]

4.15.6. The previous example takes on an even more 'paradoxical' air (for those who cannot distinguish dependence and conditional independence, or, at any rate, do not always remember that everything is relative to a given state of information) if the drawings are made without replacement.

This is the case of a 'lucky-dip': $N$ tickets are for sale (and before being sold their markings are unknown), $n$ of them are winning tickets (and one checks this by examining each ticket one has bought), which give one the right to a prize: we suppose, to avoid complications, that the prizes are identical. Conditional on the knowledge of the number of prizes, $n$, for a given number of tickets sold one's probability of buying a winning ticket is *less*, the more prizes that have been won. If, initially, one were very uncertain about the percentage of winning tickets (i.e. distributed the probability to be attributed to the various hypotheses over a wide range, for example, as a limit-case, gave equal probabilities to all the hypotheses $n = 0, 1, 2, \ldots, N$), the more frequent the occurrence of winning tickets, the more one's probability increases for the tickets yet to be sold. Under the intermediate assumption, which consists in knowing that the number $n$ has been determined by casting a die $N$ times and taking $n =$ the number of times a '6' occurs, the probability would remain constant $\left(=\frac{1}{6}\right)$ independently of any information concerning tickets sold and prizes won. (This is obvious; it is the same thing as actually playing dice: in any case, it would be a useful exercise to check the conclusion without using this direct argument.)

Examples of this kind (dice, urns, roulette etc.) are convenient because they are reduced to standard schemes. Precisely for this reason, however, they have little use or significance and, hence, it is desirable to give a more concrete and practical interpretation of the same example.

From a box containing 1000 specimens of a certain gadget, about 100 were drawn and used: 15 of them did not work properly (whereas, according to the standard specification, this should have been around five). Should one use the others or throw them away (assuming, for example, that if more than 10% were defective their use would cause more damage than the cost of throwing them away)? We shall limit ourselves to the conceptual aspects: the exact calculations, with precisely specified hypotheses, could be made now, but we shall reserve this until Chapters 11 and 12.

[15] An even better way of putting it is to say that they are 'exchangeable': we will talk about this in Chapter 11.

[16] Lindley (in the 2nd volume of *Probability and Statistics*), in order not to diverge too much from existing terminology, chose to continue to talk of *independence* (without, in cases of this kind, adding '*conditional*'). He told me that a student once objected: '*How, then, can an experience be informative*?'. This means (I observed) that your teaching is so good that it leads people to a correct understanding despite the incorrect terminology. However, it is better to use the correct terminology in order that nobody becomes confused, or has to make a strenuous mental effort in order not to be confused.

The data given say nothing except in relation to what we know, or imagine, regarding systems of production and packing. If, for packing them into boxes, the gadgets are chosen at random, there is no reason to be less (or more) confident about the remaining articles: the fact of them being together with other articles that are defective in a greater or lesser percentage is purely fortuitous. If, on the other hand, one believes that the contents of a box come from the production of a given machine at a given time, the conclusion may be different, in either sense. If one thinks that the defects are due to a machine being temporarily out of adjustment, then the usual attitude of fearing that the high percentage of defectives might also be found in the rest of the box is reasonable. If, instead, one thinks that there is a periodic cause (in an extreme case, that the seventh article in every series of 20 turns out to be defective), it is almost certain that each box contains almost exactly 50 defective pieces (at any rate, with less imprecision than under the first hypothesis). The conclusion is then the opposite one: having already removed 15 defective articles, instead of five, it is to be expected that 35 remain, rather than 45 (and the bad initial outcomes improve the prospects for the remainder, rather than making them worse).

## 4.16 Noncorrelation; Correlation (Positive or Negative)

4.16.1. The condition $\mathbf{P}(AB) = \mathbf{P}(A)\mathbf{P}(B)$ for events was referred to as both the condition for stochastic independence and the condition for noncorrelation; in the case of two random quantities, $X$ and $Y$, the same condition $\mathbf{P}(XY) = \mathbf{P}(X)\mathbf{P}(Y)$ will still be called the condition for noncorrelation (or of positive or negative correlation if either > or < is substituted for =), whereas by stochastic independence one implies a more restrictive condition, which, for the time being has only been introduced for the case of random quantities with a finite number of possible values.

One can show straightaway that the above-mentioned condition is more restrictive; in other words, that stochastic independence implies noncorrelation (but not conversely, *except in the case of two random quantities with only two possible values, and hence, in particular, for events*). Let $x_i$ ($i = 1, 2,\ldots, m'$) denote the possible values for $X$, and $p'_i = \mathbf{P}(X = x_i)$ their probabilities; similarly, let $y_j$ and $p''_j$ denote the $m''$ possible values and probabilities for $Y$. We denote the probability of the pair $(x_i, y_j)$ by $p_{ij}$; that is $P_{ij} = \mathbf{P}[(X = x_i)(Y = y_j)]$, and we observe that the $p_{ij}$, given the $p'_i$ and $p'_j$, can be any of the $m'm''$ values (lying in [0, 1]) satisfying the $m' + m'' - 1$ linear conditions $\sum_j p_{ij} = p'_i$, $\sum_i p_{ij} = p''$ (one which is superfluous, since $\sum p'_i = \sum p''_j = 1$). They are therefore determined up to

$$m'm'' - (m' + m'' - 1) = (m' - 1)(m'' - 1)$$

degrees of freedom (except in boundary cases, where some of the $p'_i$ or $p''_j$ are = 0). The condition for noncorrelation gives a further equation in the $p_{ij}$:

$$\mathbf{P}(XY) - \mathbf{P}(X)\mathbf{P}(Y) = \sum_{ij} x_i y_j \left(p_{ij} - p'_i p''_j\right) = 0,$$

which is clearly satisfied in the case of stochastic independence (we always have $p_{ij} = p'_i p''_j$),) and still allows $(m' - 1)(m'' - 1) - 1$ degrees of freedom. In other words, it permits infinitely many other solutions – that is schemes of noncorrelation without stochastic independence – unless $m' = m'' = 2$; q.e.d.

4.16.2. As for the statement that by 'strengthening' noncorrelation one can obtain stochastic independence, we were referring to the possibility of considering, besides the noncorrelation between $X$ and $Y$, the same relation between arbitrary functions of $X$ and $Y$, $X' = \alpha(X)$ and $Y' = \beta(Y)$, say: $\mathbf{P}(X'Y') = \mathbf{P}(X')\mathbf{P}(Y')$, that is $\mathbf{P}[\alpha(X)\beta(Y)] = \mathbf{P}[\alpha(X)]\mathbf{P}[\beta(Y)]$ In the case of $X$ and $Y$ with a finite number of possible values (the only case for which we have so far defined stochastic independence) it is obvious that such a relation holds, whatever the functions $\alpha$ and $\beta$ are, if $X$ and $Y$ are stochastically independent (with the above notation, if $p_{ij} = p_i' p_j''$, we have $\sum p_{ij}\alpha(x_i)\beta(y_j) = \sum p_i' p''\alpha(x_i)\beta(y_j)$. Conversely, it follows that $(m' - 1)$ $(m'' - 1) - 1$ suitable (i.e. linearly independent), additional conditions of this kind will suffice to imply stochastic independence. For the general case (an infinite number of possible values), similar conclusions will hold, except that we shall require the adjunction of *infinitely many* conditions of this kind, and, in addition, clarification of the meaning of the definition by means of suitable critical considerations (see Chapter 6).

4.16.3. If, for $X_1, X_2, \ldots, X_r$, we not only have

$$\mathbf{P}(X_i X_j) = \mathbf{P}(X_i)\mathbf{P}(X_j)$$

but also

$$\mathbf{P}(X_i X_j X_h) = \mathbf{P}(X_i)\mathbf{P}(X_j)\mathbf{P}(X_h), \text{ etc.,}$$

we could, of course, define, and look at, *noncorrelation of order three* (*or greater*) for any arbitrary distinct $X$. Equivalently (and perhaps more simply), we can say that, when $\mathbf{P}(X_i) = 0$, noncorrelation of order $k$ means that $\mathbf{P}(Z) = 0$ for each $Z$ which is the product of $h \leq k$ distinct factors $X_i$; the general case can be reduced to this one by saying that it implies noncorrelation of order $k$ of the $X_i - \mathbf{P}(\mathrm{X}_i)$. However – with a convention opposite to that for stochastic independence – when we simply say 'noncorrelation', 'pairwise' should always be understood. This is both because this is the case of most frequent interest, and in order to be able to use, in the case of events, the two convenient and easily distinguishable terms, '(stochastically) independent' and 'uncorrelated', without having to specify 'independent, that is to say, independent *of every order*' and 'uncorrelated, that is to say, *pairwise* uncorrelated', respectively.

4.16.4. Pairwise noncorrelation (unlike independence) has, in fact, an autonomous and fundamental meaning, no matter how many random quantities are being considered together. More generally, a *measure* of correlation is of interest, and this will be provided by the *correlation coefficient*, $r(X, Y)$, between two random quantities (to be defined by equation 4.24 in Section 4.16.6). In the same way as knowledge of the previsions $\mathbf{P}(X_i)$ was sufficient in order to know the prevision of every linear function of the $X_i$, $X = \sum a_i X_i$, knowledge of the prevision of the squares, $\mathbf{P}(X_i^2)$ (in addition to that of the $\mathbf{P}(X_i)$), and of the correlation coefficients $r_{ij} = r(X_i, X_j)$, is sufficient to determine the prevision of every quadratic function of the $X_i$:

$$\begin{aligned} X &= \{\text{a second} - \text{degree polynomial in the } X_i\} \\ &= \sum_{ij} a_{ij} X_i X_j + \sum_i a_i X_j + \sum_i a_i X_i + a_0,^{17} \end{aligned}$$

$$\mathbf{P}(X) = \sum_{ij} a_{ij} \mathbf{P}(X_i X_j) + \sum_i a_i \mathbf{P}(X_j) + a_0. \tag{4.20}$$

Knowledge of the *second-order previsions* is often sufficient for the solution of many problems (if not completely, by giving some bounds). If one thinks of the image (still not made precise, but intuitively clear) of *probability as distribution of mass*, the knowledge of the previsions is equivalent to the knowledge of the *barycentre*, and that of the second-order previsions (or second-degree characteristics of the distribution) is equivalent to knowledge of the *moments of inertia*.

The reasons for the importance of such knowledge, albeit limited, of the distribution in the calculus of probability (as in statistics), are, essentially, the same as those which determine their importance in mechanics (although, in general, not as precisely as is the latter case, due to the connection with energy, etc.).

4.16.5. *Separations and deviations*. It is often convenient to write

$$X = x + (X - x)$$

where $x = m = \mathbf{P}(X)$, or some other special value (like the *median* or the *mode*, which we shall discuss in Chapter 6, 6.6.6), or even with a generic $x$ (representing an arbitrary given number). We shall call the difference $X - x$ the *separation* (of $X$ from $x$); if we take the absolute value (as is often useful), $|X - x|$ is called the *deviation*.

As far as the second-order previsions are concerned, it is clear that, in general, it is convenient to take them relative to the barycentre, $x_i = m_i = \mathbf{P}(X_i)$, the point with respect to which the moments are smallest.

$$\begin{aligned} \mathbf{P}(X - x)^2 &= \mathbf{P}\big[(X - m) - (x - m)\big]^2 \\ &= \mathbf{P}(X - m)^2 + (x - m)^2 - \{2(x - m)\mathbf{P}(X - m)\}, \end{aligned}$$

but the final term vanishes ($\mathbf{P}(X - m) = m - m = 0$) and we have the following result, well known in mechanics: the moment with respect to a point $x$ is the moment about the barycentre (the first term) plus the square of the distance from the barycentre (the second term: here the mass = 1), and clearly the minimum is at $x = m$.

$\mathbf{P}(X - m)^2$ is called the *variance* of $X$, and its square root (in mechanics, the *radius of gyration*; the distance at which the mass should be concentrated in order to preserve the moment of inertia[18]) is called the *mean standard deviation* or, more briefly, the *standard deviation*. It is denoted by

[17] The first summation will suffice if we include the index 0 corresponding to the fictitious random quantity $X_0 \equiv 1$ (see Chapter 2, Section 2.8.3); in this case, $a_i$ becomes $a_{i0} + a_{0i}$ and $a_0$ becomes $a_{00}$. Moreover, it is, of course, irrelevant whether we take as zero the $a_{ij}$ with $i > j$, or conversely with $i < j$, or instead take $a_{ij} = a_{ji}$ or whatever, according to the circumstances: the only relevant thing is $a_{ij} + a_{ji}$.

[18] This is an example of a mean according to Chisini's definition! See Chapter 2, Section 2.9.2.

$$\boldsymbol{\sigma}(X) = \surd\left[\mathbf{P}(X-m)^2\right] = \surd\left[\mathbf{P}(X^2) - m^2\right] \quad \left(m = \mathbf{P}(X)\right),^{19} \tag{4.21}$$

or sometimes $\sigma_X$ (or simply o if there is no ambiguity). The variance will be denoted by $\boldsymbol{\sigma}^2(X)$, $\sigma_X^2$ or $\sigma^2$.

The separation (and the deviation) from $m$, divided by the standard deviation, are called the *standardized separation*, $(X - m)/\sigma$, and the *standardized deviation* $|X - m|/\sigma$.

In this way, we can express the square terms of Section 4.16.4 by means of previsions and variances (i.e. by means of previsions and standard deviations):

$$\mathbf{P}(X_iX_i) = \mathbf{P}\left(X_i^2\right) = \boldsymbol{\sigma}^2(X_i) + \mathbf{P}^2(X_i) = \sigma_i^2 + m_i^2 \tag{4.22}$$

$$\left(\text{where } \mathbf{P}^2(X) = \left[\mathbf{P}(X)\right]^2\right),$$

and similarly the cross-product terms, $\mathbf{P}(X_iX_j)$ with $i \neq j$;

$$\mathbf{P}\left(X_\mathrm{i}X_j\right) = m_im_j + \mathbf{P}\left[\left(X_i - m_i\right)\left(X_j - m_j\right)\right] = m_im_j + \sigma_{ij}, \tag{4.23}$$

where $\sigma_{ij}$, so defined, is called the *covariance* of $X_i$ and $X_j$,[20] and, writing $\sigma_{ij} = \sigma_i\sigma_jr_{ij}$, we arrive at the introduction of the correlation coefficient, as mentioned above.

4.16.6. In order to define the *correlation coefficient* we denote by $X$ and $Y$ the two random quantities, and suppose that $\mathbf{P}(X) = \mathbf{P}(Y) = 0$; then setting

$$\mathbf{P}(XY) = \boldsymbol{\sigma}(X)\boldsymbol{\sigma}(Y)\mathbf{r}(X,Y),$$

we have, by definition,

$$\mathbf{r}(X,Y) = \frac{\mathbf{P}(XY)}{\boldsymbol{\sigma}(X)\boldsymbol{\sigma}(Y)}. \tag{4.24}$$

It was clear from the very beginning that the correlation coefficient would be zero, positive or negative, according to whether $X$ and $Y$ are uncorrected, positively correlated, or negatively correlated. It is equally obvious that if $Y = X$, then $r = 1$, and that if $Y = -X$, then $r = -1$, and it is also clear that multiplying $X$ and/or $Y$ by constants does not change $r$, except possibly in sign:

$$\mathbf{r}(aX, bY) = \pm\mathbf{r}(X, Y),$$

+ or −, according to the sign of $ab$. If $a = 0$, or $b = 0$, then $aX = 0$ or $bY = 0$ and $r$ has no meaning; the previous observation can therefore be completed by saying that if $Y = aX$, then $\mathbf{r}(X, Y) = \pm 1$ (sign of $a$).

[19] $\sigma$ is *boldface* when it is an operator (and the same holds for $r$).

[20] In particular, for consistency, $\sigma_{ij} = \sigma_i^2$.

It is already intuitively obvious from the above that $r$ can assume all values between ±1, but no others, and we shall now prove this: it will suffice to restate the standard argument about quadratics. We always have $(Y - tX)^2 \geq 0$ (or zero, in the limit-case where for some $t = t_0$ we have the identity $Y = t_0X$), and hence $t^2X^2 - 2tXY + Y^2 \geq 0$; taking its prevision, $t^2\mathbf{P}(X^2) - 2t\mathbf{P}(XY) + \mathbf{P}(Y^2) \geq 0$, and so, since the discriminant must be negative, $|\mathbf{P}(XY)|^2 < \mathbf{P}(X^2)\mathbf{P}(Y^2)$; q.e.d.

In order to extend the definition to the case in which we do not have $\mathbf{P}(X) = \mathbf{P}(Y) = 0$, it suffices to observe that the separations from the prevision, $X - m_X$ and Y $- m_Y$, must be substituted for $X$ and Y, and $\mathbf{P}(XY)$ therefore replaced by

$$\mathbf{P}\left[\left(X - m_X\right)\left(Y - m_Y\right)\right] = \mathbf{P}\left(XY\right) - m_X m_Y.$$

It is useful to remark that a different extension of the definition could have been obtained by leaving $\mathbf{P}(XY)$ as the numerator, and changing the denominator to $\mathbf{P}_Q(X)\mathbf{P}_Q(Y)$, where $\mathbf{P}_Q(X) = \sqrt{}\mathbf{P}(X^2)$ = quadratic prevision of $X$. The same properties and proofs would hold, but the meaning would be different: if we denote this alternative coefficient (temporarily) by $\hat{r}$, $\hat{r} = 0$ would imply $\mathbf{P}(XY) = 0$, instead of $= m_X m_Y$, and $\hat{r} = \pm 1$ would follow from $Y = aX$ instead of from $Y - m_Y = a(X - m_X)$.

The meaning of all this will be clear under the geometric interpretation which we are now about to introduce.

*Remarks.* We cannot (as a rule) say that in order to have $\mathbf{P}_Q(X) = 0$ we must have $X = 0$, but only that all the probability must be at least *adherent* to 0. To have $\mathbf{P}_Q(X) = 0$, we must obviously have $\mathbf{P}(|X| \geq \varepsilon) = 0$ for all $\varepsilon > 0$ (if this were equal to $p > 0$, we would in fact have $\mathbf{P}_Q^2\left(X\right) > p\varepsilon^2$), but this does not exclude the possibility of $\mathbf{P}(X \neq 0)$ being $>0$ or even $= 1$ (e.g. if the only possible values are the sequence $x_n = 1/n$, each with zero probability). Anyway, we shall say, if $\mathbf{P}_Q(X) = 0$, that $X$ *coincides* with 0, and write $X \doteq 0$; similarly, we say that $X$ and $Y$ coincide, $X \doteq Y$, if $X - Y \doteq$ o.

## 4.17 A Geometric Interpretation

4.17.1. We have already considered (Chapter 2, 2.8.1) the linear space $\mathscr{L}$ if of random quantities $X$: it is an affine vector space (whose origin is the 'random' quantity which is identically = 0) in which each $X$ is represented by a vector (and linear combinations by linear combinations). We also agreed to denote by $X_0$ the 'random' quantity whose value is identically 1, and by $x_0$ the axis on which the 'certain' (constant) quantities lie.

Once we have introduced a prevision $\mathbf{P}$, we know that $\mathbf{P}(X)$ is a linear function of the vector $X$, with $\mathbf{P}(cX_0) = c$ (on the axis representing certainty, coinciding with the abscissa $c$). To give $\mathbf{P}$ is to give the plane of the *fair* random quantities (with $\mathbf{P}(X) = 0$): to find $\mathbf{P}(X) = m$ means, in fact, to find that $m$ for which $\mathbf{P}(X - m) = 0$; in other words, to decompose $X$ into $m + (X - m)$, the sum of a vector $mX_0$, known with certainty ($m = mX_0$), and a fair vector. One might prefer to think of $x_0 = m$ as the point of intersection of the axis of certainty with the plane parallel to the fair plane, passing through the point $X$ (where 'the point $X$' is short for $O + X$, the end point of the vector $X$ which starts from $O$).

Functions of the second degree in random quantities belonging to $\mathscr{L}$ – that is arbitrary numbers of linear combinations of products $XY$, of which the squares, $X^2$, are special cases ($Y = X$) – do not belong to $\mathscr{L}$.[21] We can, however, still give $\mathbf{P}(XY)$ a geometric interpretation by transforming $\mathscr{L}$ geometrically from an affine space into a Euclidean metric space, with a metric defined by the $\mathbf{P}(XY)$, interpreted as the *scalar product* of the vectors $X$ and $Y$*:* that is by interpreting $\mathbf{P}_Q(X)$ as the *length* of the vector $X$ (limiting ourselves to some $\mathscr{L}^* \subset \mathscr{L}$ if for $X \notin \mathscr{L}^*$ we have $\mathbf{P}_Q(X) = \infty$).

In fact, $\mathbf{P}(XY)$ satisfies the necessary and sufficient conditions for a scalar product (and therefore generates a Euclidean metric):it is linear in $X$ and $Y$, and symmetric

$$\left(XY = YX,\ X\left(Y_1 + Y_2\right) = XY_1 + XY_2,\ \mathbf{P}\text{ is linear}\right);$$

it is positive definite ($\mathbf{P}(XX) = \mathbf{P}_Q^2(X) > 0$ if we do not have $X \doteq 0$).

*Remarks.* Notice that, for the metric under consideration, it is appropriate to think of coincident random quantities as represented by the same vector (if one wished, one could say that it represents an 'equivalence class' with respect to 'coincidence'). If not, we would have nonzero vectors with zero length.

Under this metric, the length of $X$ would be $\mathbf{P}_Q(X) = \surd(m^2 + \sigma^2)$, and $X$ and $Y$ would be orthogonal if $\mathbf{P}(X\ Y) = 0$: in general, the cosine of the angle between them would be $\check{r}$. Fairness implies orthogonality to the axis of certainty. The metric that we use (most often) is not this one but another: it was, however, convenient to begin with this as it is the most natural starting point.[22]

4.17.2. The metric that serves our purpose is the same as the preceding one (in accordance with the given definition of correlation) but applied to the *separations*, $X - \mathbf{P}(X)$, instead of to the $X$ themselves. The simplest illustration (which is connected with the previous considerations) consists of saying that one takes into consideration only the projections onto the fair plane; that is the component orthogonal to the axis of certainty ($X - m$, with $m = \mathbf{P}(X)$), disregarding the parallel component, which is in fact $m$, or '$mX_0$'.

Under this metric, the length of $X$ is $\boldsymbol{\sigma}(X)$; that is the length of the projection of $X$ (under the previous metric). The cosine of the angle between $X$ and $Y$ (taking the projections onto the fair plane) is $\mathbf{r}(X, Y)$ and we have, therefore: *noncorrelation* ($r = 0$) corresponds to *orthogonality* (of the projections onto the fair hyperplane); *positive correlation* ($0 < r < 1$) and *negative correlation* ($-1 < r < 0$) correspond to *acute* and *obtuse* angles, respectively (always between the projections). The extreme cases ($r = \pm 1$) correspond to parallelism, in the same or opposite direction (again between projections).

In order to avoid constant repetition of the fact that it is the projections that are involved, one could always bear in mind that, in this ambit, if we take as norm (or length, or distance) the standard deviation instead of the quadratic prevision, all random quantities differing by certain constants are identified with one and the same vector of the fair hyperplane, the projection of the original (writing, e.g. $X \stackrel{..}{=} Y$). One must be careful not to become confused, and think in these terms when it is not possible to do so (e.g. in the case of mean-square convergence the norm must be $\mathbf{P}_Q(X)$ and not $\boldsymbol{\sigma}(X)$).

[21] They could all belong to L, if the latter were infinite dimensional; otherwise, a few of them could belong. Anyway, the appearance of $X^2$, in addition to $X^2$, is superfluous (unless one is interested in $\mathbf{P}(X^4)$, $\mathbf{P}(X^2\ Y)$, etc.).

[22] In some cases, we shall actually find it necessary to refer to the metric generated by $\mathbf{P}(X\ Y)$: e.g. in connection with *mean-square convergence* (see Chapter 6).

4.17.3. The vectorial–geometrical interpretation makes obvious and meaningful all properties relating to previsions of the second order. If we suppose that all the random quantities considered in the following are fair ($\mathbf{P}(X) = 0$), we have, for instance:

for the decomposition of $X$ into a component parallel to an arbitrary (nonzero) $Y$ and a component orthogonal, the former will be $\boldsymbol{\sigma}(X)\mathbf{r}(X, Y)$ (the length times the cosine) multiplied by the unit vector in the direction of $Y$ (i.e. $Y/\boldsymbol{\sigma}(Y)$), in other words,

$$X' = Y \cdot \left[\mathbf{r}(X,Y)\sigma(X)/\sigma(Y)\right], \tag{4.25}$$

and the latter (which is obviously $X'' = X - X'$) has length $\boldsymbol{\sigma}(X)\surd(1 - r^2)$ (length times sine). It is also characterized by the fact of having the smallest length of all vectors of the form $X - aY$;

in the same way, in order that $X'$ be contained in, and $X''$ be orthogonal to, a given linear space (for simplicity, we take it to be two-dimensional – linear combinations of $Y$ and $Z$), we will have $X' = aY + bZ$ such that

$$X'' = X - X' = X - aY - bZ$$

is orthogonal to $Y$ and $Z$; hence

$$\begin{aligned}\mathbf{P}(X''Y) &= \mathbf{P}(XY) - a\mathbf{P}\left(Y^2\right) - b\mathbf{P}(YZ) = 0,\\ \mathbf{P}(X''Z) &= \mathbf{P}(XY) - a\mathbf{P}(YZ) - b\mathbf{P}\left(Z^2\right) = 0,\end{aligned}$$

and, if $Y$ and $Z$ are taken to be orthogonal, $\mathbf{P}(YZ) = 0$, and unitary, $\mathbf{P}(Y^2) = 1 = \mathbf{P}(Z^2)$, we have straightaway

$$\begin{aligned}a &= \mathbf{P}(XY) = \sigma(X)\mathbf{r}\left(X,Y\right),\\ b &= \mathbf{P}(XZ) = \sigma(X)\mathbf{r}\left(X,Z\right),\\ X' &= \sigma(X)\left[Y\mathbf{r}\left(X,Y\right) + Z\mathbf{r}\left(X,Z\right)\right];\end{aligned}$$

with a standard procedure (similar to the above), given any linearly independent $X_1$, $X_2$, ..., $\mathrm{X}_n$, one can carry out the orthogonalization by substituting $Y_1$, $Y_2$, ..., $Y_n$, the $Y_i$ being orthogonal to each other (and, if we wish, unitary). Proceeding in order ($i = 1$, $2, \ldots, n$), it suffices to add to $X_{i+1}$ a suitable linear combination of $X_1, \ldots, \mathrm{X}_i$ in order to make it orthogonal to these vectors and, if necessary, to normalize (dividing by the length), obtaining $Y_{i+1}$;

and so on.

4.17.4. The standard deviation of the sum of two or more random quantities is particularly important. For two summands, we have

$$\begin{aligned}\sigma^2(X+Y) = \mathbf{P}(X+Y)^2 &= \mathbf{P}\left(X^2\right) + \mathbf{P}\left(Y^2\right) + 2\mathbf{P}(XY)\\ &= \sigma^2(X) + \sigma^2(Y) + 2\mathbf{r}\left(X,Y\right)\sigma(X)\sigma(Y),\end{aligned} \tag{4.26}$$

and it is easy to recognize the expression as the length of the sum of two vectors (as it had to be): that is the side of a triangle given the other two sides and the (external) angle between them; $c^2 \equiv a^2 + b^2 + 2ab\cos\theta$ (this is Carnot's theorem; if $\cos\theta = 0$, orthogonality, we have Pythagoras' theorem: in the limit cases, $\cos\theta = \pm 1$, that is parallelism, $c$ = the sum or difference of $a$ and $b$). It is important to remember the following: in the case of *orthogonality* (*noncorrelation*), the variances are added (*the standard deviations obey Pythagoras' theorem*); in the case of *positive correlation*, the variance and the standard deviation of the sum turn out to be *greater*, and in the case of *negative* correlation *less*, than in the case of noncorrelation (the standard deviations of the summands being the same) (Figure 4.2).

The same holds for more than two summands. In this case, of course, one may have correlations which are in part positive, in part negative, and the effect of either the former or the latter may prevail. The general formula is clearly as follows (written directly for a general linear form, always assuming $\mathbf{P}(X_i) = 0$):

$$\sigma^2\left(\sum_i a_i X_i\right) = \mathbf{P}\left(\sum_{ij} a_i a_j X_i X_j\right) = \sum_{ij} a_i a_j \mathbf{P}\left(X_i X_j\right) = \sum_{ij} a_i a_j \sigma_i \sigma_j r_{ij}; \tag{4.27}$$

the squared terms ($r_{ij} = 1$) yield $\sum_i a_i^2 a_i^2$; excluding $i = j$ in the general summation, one obtains the contribution of the cross-product terms (zero in the case of orthogonality, positive or negative according to the prevailing correlations *between the summands* $a_i X_i$ – not the $X_i$ ! – whose signs are those of $a_i a_j r_{ij}$ – not of $r_{ij}$!).

The *covariance matrix*, with entries $\sigma_{ij}$, of the random quantities $X_i$ (which we assume to have zero prevision) completely determines the second-order characteristics in the space $\mathscr{L}$ of linear combinations of the $X_i$ (geometrically, in $\mathscr{L}$, it gives the length and angles of the vectors representing the $X_i$). The *correlation matrix*, with entries $r_{ij}$($r_{ij} = \sigma_{ij}/\sigma_i\sigma_j$, $\sigma_i = \sqrt{\sigma_{ii}}$, $r_{ii} = 1$) can be derived from it, giving the angles ($r_{ij}$ is the cosine) but not the lengths. It can still be regarded as a covariance matrix for the standardized $X_i$; that is for the $X_i/\sigma_i$ (geometrically one is considering the *unit vectors* rather than the vectors).

4.17.5. A fact that is of conceptual and practical importance – and for this reason mentioned already in the Remarks in Section 4.9.1. for the case of events – is that the size of the *negative* correlation (unlike the positive) must be *bounded*. More precisely, given $n$ random quantities, the arithmetic mean of their $\binom{n}{2}$ correlation coefficients $r_{ij}$ ($i \neq j$) cannot be less than $-1/(n-1)$: in particular, the $r_{ij}$ cannot all be less than $-1/(n-1)$; in the extreme case (as we shall see) they can all be equal to this limit value.

Without loss of generality, we can assume the $X_i$ normalized, $\mathbf{P}(X_i) = 0$ and $\mathbf{P}(X_i^2) = 1$, so that $r_{ij} = \mathbf{P}(X_i Y_j)$: we consider their sum, $X = X_1 + X_2 + \ldots + X_n$, and evaluate its variance

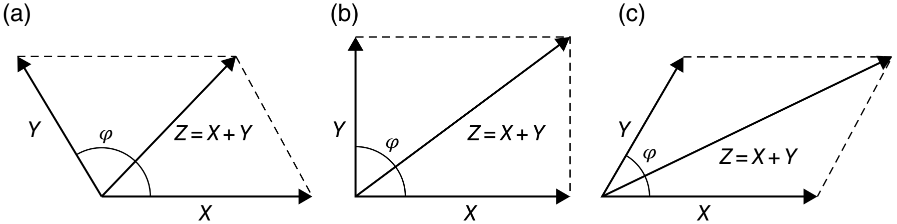

**Figure 4.2** (a) Negative correlation. (b) Noncorrelation (orthogonality). (c) Positive correlation.

$$\sigma^2(X)=\mathbf{P}(X^2)=\mathbf{P}\left(\sum_{ij}X_iX_j\right)=\sum_{ij}\mathbf{P}(X_iX_j)=\sum_{i}\mathbf{P}(X_i^2)+\sum_{i\neq j}\mathbf{P}(X_iX_j)$$
$$=n+\sum_{i\neq j}r_{ij}=n+n(n-1)\bar{r}=n\left[1+(n-1)\bar{r}\right],$$

where we have set $\bar{r}$ = the arithmetic mean of the $r_{ij}$, that is

$$\bar{r}=\frac{1}{n(n-1)}\sum_{i\neq j}r_{ij}.$$

The variance is non-negative, however, and therefore $\bar{r} \geq -1(n-1)$; q.e.d. We note that the extreme value is attained if and only if the sum is identically = 0 (or, if we want to be absolutely precise, $\doteqdot 0$, using the notation of Section 4.17.2): that is if the $n$ unit vectors have zero resultant.[23] In particular, the $r_{ij}$ could have the common value $r = -1/(n-1)$ only if the unit vectors were arranged like the straight lines joining the centre of a regular ($n$ – l)-dimensional simplex to the vertices. Figure 4.3 illustrates the case of $n$ = 3 (equilateral triangle) and $n$ = 4 (regular tetrahedron). We give the basic facts for these cases (and also for $n$ = 5, 6, 7, 8):

$$n=3, r=-1/2=\cos 120^\circ \qquad n=6, r=-1/5=\cos 101^\circ 32'$$
$$n=4, r=-1/3=\cos 108^\circ\ \ 16' \qquad n=7, r=-1/6=\cos 99^\circ 36'$$
$$n=5, r=-1/4=\cos 104^\circ 29' \qquad n=8, r=-1/7=\cos 98^\circ 12'$$

Approximately, the angle is a right angle plus l/($n$ – 1) (in radians); in other words, in a possibly more convenient form, plus 3438/($n$ – 1) minutes (for $n$ = 8 the error is already of the order of 1′). These numerical examples serve to make clear that one cannot go much beyond orthogonality among random quantities when there are more than just a few of them.

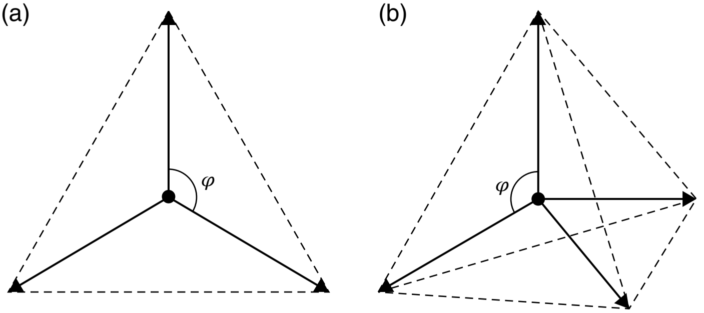

**Figure 4.3** (a) The maximum negative correlation for three vectors: $r=\cos\phi=-\frac{1}{2}$. (b) The maximum negative correlation for four vectors: $r=\cos\phi=-\frac{1}{3}$.

[23] Observe that they are, therefore, linearly dependent.

4.17.6. All that we have considered so far (Sections 4.17.2–4.17.5) has been in terms of the conventional representation of the $X_i$ (and of the $X$ linearly dependent on them) in the abstract space $\mathscr{L}$. If, instead, we wish to consider the meaningful interpretation in terms of the distribution of probability as distribution of mass – an interpretation whose importance was indicated at the end of Section 4.16.4 – we must transfer to the linear ambit $\mathscr{A}$ (the space $S_r$, with coordinates $x_1$, $x_2$, $x_r$, where a point represents the outcomes of $X_1, X_2, \ldots, X_r$), since it is over this space that the mass is distributed. The $\mathbf{P}(X_i) = x_i$ identify the barycentres of such distributions (and we again assume the barycentre coincident with the origin, in order to avoid useless petty complications in the notation), and the $\mathbf{P}(X_iX_j) = \sigma_{ij}$ identify the moments of inertia; that is *the ellipsoid* (or *kernel*) *of inertia* (and in our case it could be called *of covariance*, like the corresponding matrix).

For our purposes, it is much more meaningful and useful (although the two things are formally equivalent) to consider what we shall call the *ellipsoid of representation*,[24] which is the reciprocal of the other. With reference to the principal axes (common to the two ellipsoids), the semi-axes measure the corresponding standard deviations, $\sigma_h$, in the ellipsoid of representation, whereas, for the ellipsoid of covariance, they give the reciprocals, $1/\sigma_h$ (or $K/\sigma_h$; one can take an arbitrary multiplicative constant).

In Mechanics, the latter has been employed (Cauchy–Poinsot), although the former has also been proposed (MacCullach). Part of the reason for preferring this one seems also to hold for Mechanics; in our case, however, there are also rather special and more decisive circumstances (e.g. the fact that we are interested in moments with respect to planes, that is, in general, to hyperplanes $S_{r-1}$, rather than moments with respect to straight lines).

The ellipsoid of representation has a concrete meaning: it is the model of a solid having the same moments as the given distribution (assuming it to be homogeneous, and giving it a mass increased in the ratio 1 to $r + 2$ – three on the line, four in the plane, five in ordinary space and so on – or, alternatively, increasing the size in the linear scale 1 to $\sqrt{(r + 2)}$). This is obvious if one thinks of the case of the sphere, to which one can always reduce the problem by imposing a suitable metric on the affine space $\mathscr{A}$ (unless it already has one, either because of an actual geometrical meaning, or because the arbitrariness has already been exploited by reducing to a sphere some ellipsoid previously considered). For the unit sphere (in $S_r$) the moment about the centre is

$$\int_0^1 \rho^2 \rho^{r-1} \mathrm{d}\rho / \int_0^1 \rho^{r-1} \mathrm{d}\rho = r/(r+2),$$

but it is also $r$ times the moment about a diametrical hyperplane, and hence the latter is $l/(r + 2)$. In order to make this equal to 1, it is sufficient to increase either the mass or the radius in the above mentioned way.

In the case of probability and statistics, this reduction to a homogeneous distribution is not the most appropriate procedure: the standard example of the ($r$-dimensional)

[24] Of course, we speak of 'ellipsoids' in $S_r$, even if $r > 3$, or $r = 2$ (ellipses), or $r = 1$ (segments). As far as I know, terminology of this kind does not exist in mechanics; statisticians at times refer to the 'ellipsoid of concentration'.

normal distribution is much more meaningful (well- known as the 'distribution of errors'). As we shall see when we come to discuss it (Chapter 7, 7.6.7 and Chapter 10, 10.2.4), to each distribution over $S_r$ there corresponds a unique normal distribution having the same second-order characteristics (same covariance matrix), and the ellipsoid of representation characterizes it in the most directly expressive manner.

These brief comments may have led to an appreciation of how many interesting conclusions, although incomplete, of course, can be drawn from incomplete assumptions (even as incomplete and crude as in the case under consideration).

4.17.7. *Inequalities*. We must now establish certain inequalities that are both necessary for the topic in hand and also serve as simple illustrations of what can be said more generally.[25]

*Tchebychev's inequality* gives an upper bound, $1/t^2$, for the probability that $|X|$ is greater than $t\mathbf{P}_Q(X)$; in particular, *for the probability that the standardized deviation is greater than t.* For example, the probability that $|X|$ is greater than some multiple of the quadratic prevision is: $< \frac{1}{4}$ for twice; $< \frac{1}{9}$ for three times; $< \frac{1}{25}$ for five times; $< \frac{1}{100}$ for ten times and so on. Without further conditions, this bound is the best possible; however, the bounds are normally crude (the probability is much smaller: we have here placed ourselves in the least favourable position).

The *proof* is obvious if one thinks in terms of mass. If a mass $>1/t^2$ were placed at a distance from the origin $>a$, it would have moment of inertia $>a^2/t^2$; altogether, the moment of inertia is $\mathrm{P}_Q^2(X)$ and hence $a < t\mathbf{P}_Q(X)$. Placing two masses $1/2t^2$ at $\pm\, t\mathbf{P}_Q(X)$ and the rest at 0, one obtains the limit-case (provided that $t \geq 1$).

*Cantelli's inequality* is the one-sided analogue of the preceding one: $1/(1 + t^2)$ is the upper bound for the probability that the separation *in a given direction* is greater than $t\sigma$ ($X > m + t\sigma$, or $X < m - t\sigma$, respectively, with $t > 0$). If the mean is not fixed, the question does not arise, the inequality would then be the same as the first one; the improvement is notable only for small $t$: $t = \frac{1}{2}, p = \frac{4}{5}$ instead of 1; $t = \frac{3}{4}, p = \frac{64}{100}$ instead of 1; $t = 1, p = \frac{1}{2}$ instead of 1; $t = \frac{3}{2}, p = \frac{4}{13}$ instead of $\frac{4}{9}$; $t = 2, p = \frac{1}{5}$ instead of $\frac{1}{4}$; for $t = 3$ the difference is already hardly noticeable: $p = \frac{1}{10}$ instead of $\frac{1}{9}$

The proof can be given in a similar way to the above. In order to balance a mass $p$ at $m + t\sigma$, one can place the residual mass $1 - p$ at $m - t\sigma p/(1 - p)$, and this gives a moment of inertia equal to $\sigma^2 t^2[p + (1 - p)p^2/(1 - p)^2]$; $t^2[\ldots]$ cannot be greater than 1, $[\ldots] = p/(1 - p)$, $t^2 \leq (1 - p)/p = -1 + 1/p$ and so on. If the balancing mass is dispersed, the situation can only be made worse.

Although it is outside of our present realm of interest (second-order characteristics), it is worthwhile pointing out how the argument used in proving Tchebychev's inequality can be applied, without any difficulty, to much more general cases. If $\gamma(x)$ is an increasing function ($0 \leq x \leq \infty$), we necessarily have $\mathbf{P}\{|X - m| \geq a\} \leq \mathbf{P}\{\gamma(|X - m|)\}/\gamma(a)$ because a mass $>p$, placed at a distance $a$ from $m$, alone contributes to $\mathbf{P}\{\gamma(|X - m|)\}$ a quantity $>p\gamma(a)$ (which cannot be greater than the whole thing), and the situation is even worse if the distance is greater.

[25] More general cases than those considered here are developed in the works of E. Volpe (using this geometrical representation): Ernesto Volpe di Prignano, 'Calcolo di limitazioni di probabilité mediante involucri convessi', *Pubbl. n. 16 dell'Ist. Matern. Finanz. Univ. di Trieste* (1966).

For example, taking absolute moments of any order $r$, we have

$$\mathbf{P}\Big(|X|\geqslant a\Big)\leqslant\mathbf{P}\Big(|X|^r\Big)|\,a^r,$$

the Markov inequality: for $r = 2$ this is the Tchebychev case, seen above.

## 4.18 On the Comparability of Zero Probabilities

4.18.1. When we were considering (at the end of Chapter 3) countable additivity and zero probabilities, the question often arose as to whether it makes sense to compare the latter; for example, saying that, if all cases are equally probable, the probability of the union of 12 of them is twice that of the union of six, and three times that of the union of four, even if all these probabilities are zero (as in the example of 'an integer $N$ chosen at random'). We assumed this in order to give the statements of a few examples in a more suggestive form; as we indicated then, this is now the time to examine the question.

For the purpose of removing the most radical objection, and as a better means of presenting the sense of the question, a geometrical analogy will suffice. The objection is that zero stands for nothing, and that nothing is simply nothing: this is one of many such vacuous statements on the basis of which certain philosophers pontificate about things of which they understand nothing.[26]

A set can have measure zero in terms of volume without being empty; it could, for instance, be a part of a surface and have a measure in terms of area (and two areas can be compared). A measure in terms of area could be zero without the set being empty; it could be an arc of a curve and have a measure in terms of length. A linear set might also have measure zero in terms of length (in some sense or other: Jordan–Peano, Borel, Lebesgue) without being empty, but some comparison could also be made in this case (even if it only distinguished sets with single points or 2 or 3,..., or an infinite number).

All this would be even more expressive and persuasive if put in terms of more general concepts of measure (with intermediate dimensions also, not just integer) as in Borchardt, Minkowski, Peano, Hausdorff and so on. The example closest to our theme is that in which one defines 'the measure $m$ of dimension $\alpha$' of a set $\boldsymbol{I}$ to be that for which $\boldsymbol{V}(I_\rho) \sim m\rho^{3-\alpha}$ ($I_\rho$ = the set of points of three-dimensional space with distance $\leqslant \rho$ from $I$, $V$ = volume, the asymptotic expression to hold as $\rho \to 0$).

4.18.2. In any case, so far as probability is concerned, a direct meaning exists and we have no need of analogies to provide a justification (they may, on occasion, provide encouragement in showing us that our situation is not unique and strange, and may help us by providing visually intuitive models).

Given two events $A$ and $B$, it is clear that if one has to decide between them – that is if one makes the assumption that one of the two is true – a comparison of their

[26] *Translators' note*. The author is here referring to what he considers the deleterious influence of Croce's idealism upon Italian culture.

probabilities must be made. Expressed mathematically, if we consider their probabilities $\mathbf{P}(A|H)$, $\mathbf{P}(B|H)$ conditional on the 'hypothesis' $H = A \vee B$, their sum is $\geq 1$, and their comparison is easy. It could be said that this is the same thing as comparing $\mathbf{P}(A)$ and $\mathbf{P}(B)$: if $\mathbf{P}(H)$, and, *a fortiori*, $\mathbf{P}(A)$ and $\mathbf{P}(B)$, are small, however, the proposed alternative is perhaps psychologically more appropriate as it presumably induces one to weigh up the evaluation more accurately by fixing attention on the two cases separately, whereas the reliability of the ratio of two very small numbers – attributed as part of an overall evaluation, in which $A$ and $B$ had no special significance – might well be doubted. When the events $A$ and $B$ (and hence $H$) have zero probabilities, however, the alternative approach becomes essential. With the direct comparison the ratio of the two probabilities would have the form 0/0. This does not mean that the ratio is meaningless, but that the method of comparison is not the right one.[27]

From an axiomatic viewpoint, the extension of the condition of coherence to cover the present case requires a stronger form: we assume tacitly that this has been done (but we will discuss it in the Appendix, Section 16).

Hence, with any event $A$ as reference point, any other event $E$ has a certain ratio of probability with $A$ (a finite positive number, or zero, or infinity): in this way, innumerable 'layers' of events having probabilities 'of the same order' (that is with finite ratio) can appear, the 'layers' being ordered in such a way that every event in a higher layer has infinitely greater probability than any event in a lower layer.

4.18.3. An example will suffice as a clarification, both of the general situation, and of the implicit applications mentioned in Chapter 3: this is the example of a 'positive integer $N$ chosen at random'.

We have a partition into an infinite number of events, $E_h = (N = h)$, all with zero probabilities, $\mathbf{P}(E_h) = 0$ $(h = 1, 2, \ldots)$. This says very little, however; it merely excludes a single case $(\Sigma_h p_h > 0)$ which, from this viewpoint, is 'pathological' (in the sense that, if we think of a function as having been chosen among the entire, unrestricted class of functions of a real variable, to be continuous, even at a single point, is a pathological case). To say that 'all the events $E_h$ are equally probable' is a rather substantial addition: nevertheless, it only suffices to enable us to conclude the following: if $A$ and $B$ are finite unions of the $E_h$, for example of $m$ and $n$, respectively, then the ratio of their probabilities is $m/n$; if $A$ is the complement of a finite set we certainly have $\mathbf{P}(A) = 1$; if $A$ and its complement are infinite, then $\mathbf{P}(A)$ is infinitely greater than any of the $\mathbf{P}(E_h)$, but can be any $p \geq 0$ (even $p = 1$, or $p = 0$) located somewhere in the scale of the 'layers'.

At first sight, it might seem that one could say something more (perhaps by considering frequencies for the first $n$ numbers and then passing to the limit): for example that the probability of obtaining $N$ even is $= \frac{1}{2}$, of obtaining $N$ prime is $= 0$, nonprime $= 1$. In fact, this is not a consequence of the assumption of equiprobability at all; it is sufficient to observe that, by altering the order, these limits change but the equiprobability does

[27] The knowledge that on a day when a housewife has not bought any sugar she has spent 0, does not allow us to conclude that the price of sugar is meaningless because it is 0/0; it merely indicates that the information available is not sufficient to determine it.

not; on the other hand, the possible evaluations are not only those of the limit-frequency type, up to rearrangements.[28]

The assumption that $\mathbf{P}(E) = \lim \mathbf{P}(E|N \leqslant n)$ (and possibly, more generally, $\mathbf{P}(A)/\mathbf{P}(B) = \lim [\mathbf{P}(A|N \leqslant n)/\mathbf{P}(B|N \leqslant n)]$; i.e. the limit of the ratio of the numbers of occurrences of $A$ to those of $B$ in the first n integers) is neither compulsory nor ruled out (for any $E$, or pairs $A$, $B$) where the limit exists. One certainly obtains a coherent evaluation (by continuity; see Chapter 3, 3.13) in the field where the limit exists, extendable everywhere (Chapter 3, 3.10.7). However, one makes the arbitrary choice from among the infinite possible ones, and automatically satisfying the conditions $\lim\inf \mathbf{P}(E|N \leqslant n) \leqslant \mathbf{P}(E) \leqslant \lim\sup \mathbf{P}(E|N \leqslant n)$.

This choice has no special status from a logical standpoint but it could be so from a psychological point of view if the order has some significance (e.g. chronological); and indeed it is so if the formulation in terms of an infinite number of possible cases is thought of as, more or less, an idealization of the asymptotic study of the finite problem, with a very large number of cases $n$.

One can observe, by means of this example, just how rich the 'scale' of layers' can be (perhaps more than one would imagine at first sight). For every function $\phi(n)$, tending to zero as $n \to \infty$, we can construct an event (a sequence of integers, $a_1 < a_2 < \ldots < a_n, \ldots$) in such a way that the frequency $(n/a_n)$ tends to zero like $\phi(n)$. It is sufficient to insert into the sequence, as the term $a_{n+l}$, the number $m$ if otherwise $n/m$ would be less than $\phi(m)$. If we consider $\phi(n) = n^{\tilde{}\alpha}$ $(\alpha > 0)$, we obtain, for example, an event $E_\alpha$, and each $E_\alpha$ has infinitely greater probability than those with a larger $\alpha$ (and, as is well known, the scale is far from being complete: one could insert the $E_{\alpha, \beta}$ corresponding to $\phi(n) = n^{-\alpha}(\log n)^\beta$; and so on).

4.18.4. The method of taking limits, either starting from finite partitions (e.g. $p_h^{(n)} = 1/n$ for $h = 1, 2,\ldots, n$), or countably additive ones (e.g. $p_h^{(n)} = Ka^h$, $a = 1 - 1/n$, $K = n^2/(n - 1)$, $h = 1, 2, \ldots$), with limits which are not countably additive, is, in any case, the most convenient way of constructing distributions that are not countably additive. We must bear in mind, however, that it is a procedure for obtaining *some* coherent distributions in the field in which they are defined by the passage to the limit (since *finite* additivity is preserved), and not necessarily a procedure expressing anything significant.

In particular, one should not think (even inadvertently):

> that, assuming the $p_h^{(n)}$ are probabilities conditional on an hypothesis $H_n$ (e.g. $N \leq n$, in the first example), the $p_h = \lim p_h^{(n)}$ (and the distribution over infinite subsets which derives from these) give probabilities that are conditional on the hypothesis $H = \lim H_n$ (e.g. referring still to the first example, $H = 1$);

or, even worse, the converse;or that the events for which probabilities are defined by virtue of the passage to the limit have any special rôle, or that their probabilities have a

[28] If while progressively attributing probability to infinite subsets of events (as in Chapter 3, Section 3.10.7) we always attribute probability = 1 (provided it is not necessarily = 0 by virtue of previous choices), we obtain an *ultrafilter* of events with probability = 1, whereas all the others have probability = 0. Linear combinations of distributions of this 'ultrafilter type' form a much wider class, still disjoint, however, from those of the limit-frequency type.

different meaning from those of the other events (apart from the trivial observation that the former are consequences of the evaluations made by deciding to base oneself, on the passage to the limit, whereas the latter require a separate evaluation: it could have been the other way around if we had started with a different procedure).

4.18.5. Procedures of this kind have often been employed, more or less as a result of interpretations of the type we have here rejected. The most systematic treatments known to me are those by A. Lomnicki (*Fundamenta Mathematicae*, 1923) and by A. Rényi (in many recent works; see, for example, *Ann. Inst. Poincaré* (1964): prior to this, in German, 1954).

Rényi's approach is constructed with the aim of making considerations of initial probabilities for partitions which are not countably additive fall within the range of the usual formulations, by concealing the nonadditivity by means of the passage to the limit. The device consists in accepting that, for the partitions under consideration, countable additivity must be respected, but, in the passage to the limit, the total probability may become *infinite* instead of *one*. The importance of this is mainly in connection with the inductive argument, so we will return to this topic more explicitly in Chapter 11.

## 4.19 On the Validity of the Conglomerative Property

4.19.1. If, conditional on every event $H_j$ of a finite partition, the probability $\mathbf{P}(E|H_j)$ of a given event $E$ is $p$ (or, respectively, lies between $p'$ and $p''$), then we also have $\mathbf{P}(E) = p$ (or, respectively, $\mathbf{P}(E)$ lies between $p'$ and $p''$). In fact, we have

$$\mathbf{P}(E)=\mathbf{P}(EH_1+EH_2+\ldots+EH_n)=\sum_j\mathbf{P}(E|H_j)\mathbf{P}(H_j)=p\sum_j\mathbf{P}(H_j)=p; \tag{4.28}$$

the same holds even if the $H_j$ form an infinite partition, so long as the sum of their probabilities is = 1. In fact, if we put

$$H_n^*=1-(H_1+H_2+\ldots+H_n),$$

we have

$$\mathbf{P}(E)=\sum_j\mathbf{P}(E\mid H_j)\mathbf{P}(H_j)(j\leqslant n)+\mathbf{P}(EH_n^*)=p\left[1-\mathbf{P}(H_n^*)\right]+\mathbf{P}(EH_n^*), \tag{4.29}$$

and hence $\mathbf{P}(E) = p$ because $\mathbf{P}(H_n^*)$ and, *a fortiori*, $\mathbf{P}(EH_n^*)$ tends to 0 as $n$ increases.

4.19.2. Indeed, it would appear natural that this (conglomerative) property should hold for logical reasons, overriding all mathematical demonstrations or justifications, especially if one interprets literally a phrase like 'conditional on each of the possible hypotheses the probability of $E$ is $p$, and so the fact that $\mathbf{P}(E) = p$ is proved'.

Two counterexamples will demonstrate that this is not so.

Taking an infinite partition of the integers into finite classes (each of three elements) we consider the events $A_h = E_h + E_{2h} + E_{2h+2}$, with $h = 1, 3, 5,\ldots$ odd; conditional on each of the $A_h$, the probability that $N$ be even is $\frac{2}{3}$; the analogous partition $B_h = E_{h+1} + E_{2h-1} + E_{2h+1}$ would instead give $\frac{1}{3}$ (the asymptotic evaluation gives $\frac{1}{2}$).

Consider an infinite partition of the integers into infinite classes, with $A_h$ ($h$ odd) containing the number $h$ and all multiples of $2^h$ which are not multiples of $2^{h+2}$; conditional on every $A_h$, the probability that $N$ be even is $= 1$ (independently of any conventions like asymptotic evaluations, there is only one odd number versus an infinite number of even ones and they are all equally probable). Of course, it suffices to change $N$ into $N + 1$ in order to obtain the opposite conclusion: the probability that $N =$ even is 0 conditional on every $A_h$.

When we are in a position to discuss independence and dependence for general random quantities (Chapter 6, 6.9.5; see also Chapter 12, 12.4.3), we shall meet an example which is more meaningful, both from an intuitive and practical point of view (the latitude and longitude of a point of the earth's surface 'chosen at random').